AG DANK Car Configuration Dataset 2018

Marvin Schweizer & Andreas Geyer-Schulz

September 5, 2018

The *AG DANK Car Configuration Dataset 2018* originates from the online product configurator of a major German car manufacturer. This dataset contains nearly a million configurations of about half a million users. It is likely one of the largest datasets of its kind and may be of considerable value for research in fields such as economic choice analysis, mass customization and marketing analytics.

Product or service configurators are software-based expert-systems that support their users in specifying the configurable components, called *attributes*, of a given product or service [Mittal and Frayman, 1989]. The user configures each attribute, by choosing one option from a set of possible options of the respective attribute. The options of an attribute are called *attribute levels*.

The dataset contains the configurations made by real users of this company's configurator, referred to as *respondents*.

The original dataset was published by Peter Kurz from TNS Infratest for the workshop of the special interest group for data analysis of the German Classification Society (AG DANK) in Karlsruhe on 20.-21. November 2015. The following dataset description is based on an excerpt from [Schweizer, 2018, chapters 2 & 3].

1 The Dataset at a Glance

This section provides a high-level description of this dataset and should enable the reader to start exploring the dataset quickly. All available information that is included and subsequently described is summarized in table 1.

1.1 Choice Data

The following relational model with the relations `respondents`, `configuration_types`, `configuration_prices` and `choices` contains these choices for a sample of respondents:

- `respondents` (respondent_id, budget)
  Each respondent's budget. In the current version of the dataset, this is not available. For an explanation of this absence, see section 2.

- `configuration_types` (type_id, first_attribute,..., last_attribute)
  The specification of each configuration type based on the configured attribute levels. Wherever available, plain text names have been used instead of encodings for attribute names and attribute levels.

  Due to this relation's large number of fields, their description has been abbreviated above. Especially the field names *first_attribute* and *last_attribute* only function as an informal description of what this relation looks like. Table 4 describes all fields of this relation in detail.

  See section 3 for a description of the configuration type concept and the appendix for more details on the configuration attributes.

- `configuration_prices` (type_id, sequence_number, price)
  All available prices for configurations. *sequence_number* is an integer between 1 and 4 that identifies if this is the respondent's first, second, third or fourth configuration. Configuration prices of otherwise identical configurations vary across configuration sequence numbers as described in the appendix.

- `choices` (respondent_id, type_id, sequence_number)
  The recorded configuration choices for all respondents, referencing the respondent and configuration type by the respective foreign keys.

The underlining indicates primary keys of the respective relation. All keys are integers.

1.2 Additional Information

In the following, we will consider some additional meta-information that is included in the dataset. Namely, plain text names for most attribute- and attribute-level-encodings, some price information for attribute levels as well some information about the data generation process:

- The mapping between attribute (level) codes and attribute (level) names in English is provided in the files `attributes.csv` and `attribute_levels.csv`, respectively. We substituted plain text names for encodings in the dataset wherever possible, to improve readability. Such mapping was not available for the engine attribute. This attribute is therefore not included in the respective file. In the `attribute_levels.csv` file, the boolean-valued attributes are also not included.

- The available price information is provided in the file `attribute_level_prices.csv`. Note that price information was generally unavailable for the attributes engine, line and accessories. These attributes are therefore missing from the file altogether. For attributes where prices where generally available but missing in a specific case, the respective field was left blank.

- The available information about the data generation process is provided in section 4 and in `configuration_steps.csv`. Note that the 36 binary attributes that constitute *accessories*, are configured in the same configuration step (number 7) while all other attributes have their "own" configuration step.

Note that, since the plain text attribute (level) names have been substituted for the numeric encodings in the provided dataset, the above mappings are not necessary in order to work with the dataset. However, when necessary, these mappings can be used to encode the dataset for special types of analysis.

Information entity	**Where to find it**	**Comment**
Respondent configuration choices	`choices.csv`	-
Respondent budgets	`respondents.csv`	Budget not available
Configuration types	`configuration_types.csv`	See table 4 for a specification
Configuration prices	`configuration_prices.csv`	-
Attribute level prices	`attribute_level_prices.csv`	See also table 2
Attribute encodings	`attributes.csv`	See also tables 2 and 3
Attribute level encodings	`attribute_levels.csv`	See also table 2
Configuration process	`configuration_steps.csv`	See also figure 1

Table 1: Overview of the information contained in this dataset and where to find it.

2 Dataset History

2.1 AG DANK Dataset 2015

Peter Kurz from TNS Infratest published a dataset of car configurations for the AG DANK workshop 2015 in Karlsruhe. It is a sample of a larger, unpublished dataset that includes up to nine car configurations from about 30 million respondents that were recorded in the first half of 2012.

The sample was obtained by selecting the data of three randomly drawn days from this period in 2012 and contains a total of 962799 configurations from 469112 respondents. It has been limited to the first four out of nine configurations of each respondent, since only a negligible number of respondents have configured more than four cars.

2.2 AG DANK Dataset 2018

The present dataset resulted from further processing the dataset published for AG DANK 2015 for the usage in subsequent research efforts ([Fuhrmann et al., 2016, Fuhrmann et al., 2017, Schweizer, 2018, Rüde et al., 2016]):

- About 6% of the configuration records were broken and could only be recovered partly. After cleansing, the dataset included 929576 configurations of (still) 469112 respondents, all of which configured at least one and up to four cars.
- The available budget of each respondent was originally included in the dataset but has been discarded because it had not been used in the above mentioned research projects. In the present dataset it is left blank.
- Without any information loss other than the (intentionally) discarded budget information, the size of the dataset has been reduced by an order of magnitude through normalization of the data model.
- (Meta-)information that was spread across a number of unstructured files has been made more easily accessible through a number of parsing, interpretation and consolidation steps. In this regard, names and descriptions have also been translated to English.

3 Lossless Compression by Normalization

A remarkable feature of the dataset is the high concentration within the vast configuration space: Out of $1.25 \cdot 10^{17}$ possible configurations only 943 different configurations occur in this fairly sized dataset. Each of these 943 configurations occurs as some user's first configuration. The number of different configurations in the set of the second, third and fourth configurations declines to 711, 490, and 262, respectively.

As described in [Fuhrmann et al., 2016], we can exploit this feature to reduce the size of the dataset by an order of magnitude through normalization. For this normalization, we think of identical configurations of different respondents as being of the same *abstract configuration type* (or just *type*). In the data model, each respondent entity is now merely referencing up to four of these configuration types (one for each configuration made by the respondent), thus achieving the normalized model with the four relations `respondents`, `configuration_types`, `configuration_prices` and `choices`, described above.

4 Data Generation Process

The car configurations (and additional information) contained in this dataset originate from an online car configurator. This particular type of product configurator is used by virtually all car manufacturers, including Ford, Volkswagen, Toyota, Renault and BMW (cf. [cyLEDGE Media,

2018]). Most of these configurators have substantial overlaps in the structuring of the configuration procedure and the general functionality.

These configurators guide the user by outlining a configuration process that consists of about 4-7 steps. (Roughly one for each group of attributes.) The user is encouraged to follow this ordering of the step-by-step process by the various design elements of the respective configurator. However, all of these configurators have in common that users can move back and forth between the configuration steps and even disregard the configuration step of any attribute. Default attribute levels are provided for each attribute. (If a user chooses not to configure an attribute, the default level is configured.) Furthermore, a configurator notifies the user immediately if the desired configuration violates a constraint. Such constraints may be some kind of incompatibility constraints that prohibit a particular combination of attribute levels. For example, sports seats combined with a luxury package or two exterior colors. They may also represent preconditions which state that a certain attribute (set) is a precondition for configuring another attribute. For example, a sports package may be a condition for choosing a special high-performance breaking system.

The fact that no information about such constraints is provided with the dataset makes the data generation process - at least to some degree - a black box. One has to take this into account, e.g. when analyzing the absence or presence of certain attribute combinations. Because of this, calculating the exact number of possible combinations is not possible, too. Only the upper bound of approximately $1.25 \cdot 10^{17}$ can be given by assuming the total absence of constraints.

In the case of a constraint violation, a configurator proposes different actions in order to attain an admissible configuration. The configurators only differ marginally in the way they group the configurable attributes or in their ordering of the configuration steps.

The insights obtained from analyzing the general structure of configurators in the automotive industry combined with evidence obtained from the original dataset - specifically, the dataset description and the naming of the car configuration attributes - leads us to the presumable structure of the configuration process as depicted in figure 1.

The consumer first specifies the six attributes *Engine*, *Line*, *Color*, *Rims*, *Upholstery* and *Trims*, in that order, before configuring the 36 binary attributes. Note, that we have combined these binary attributes in a single process step even though they are distinct attributes which are not mutually exclusive.

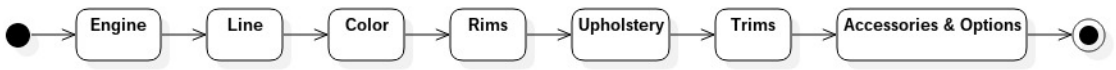

Figure 1: UML 2.2 Activity Diagram of the expected car configuration process.

The dataset included no purchase information nor the information about whether the user had the chance to order the configured car in general.

5 Questions

Potential questions that could guide an analysis of this dataset were included in the original dataset description by Peter Kurz and are attached below. Of course, other interesting approaches to analysing the data are also welcome.

- Which *Accessories & Options* items could be bundled? Bundles can reduce variation in configurations and thus result in economies of scale.
- What levels of the attribute *line* can function as substitute for one another? The manufacturer could then reduce the number of lines.
- Does a user have purchase intention or is she only using the configurator as a pastime?

Appendix: Car Attributes and `configuration_type` Relation

Each car configuration is described by 42 attributes and a total configuration price. The total configuration price is only provided for each user's first configuration. Six attributes are non-binary - meaning that they have more than two possible levels. These six attributes are *Engine* (9 levels), *Line* (4), *Color* (12), *Rims* (24), *Upholstery* (16) and *Trims* (11).

For each of these attributes (except for *Engine*) a human understandable name for each level is given. For example, the levels of the *Line* attribute are *No Line*, *Sports Line*, *Luxury Line* and *Modern Line*. For the levels of *Engine* only numerical identifiers (1,2,...,9) were provided in order to keep the manufacturer and car model undisclosed.

For each level of these attributes, a price information is provided, except for *Engine* and *Line*. The price for a distinct attribute level differs depending on the configuration number. For example, the price of the *Color* level *Black Sapphire Metallic* is 840€ when configured in a first configuration and 1008€ when configured in a fourth configuration.

The remaining 36 attributes are *Accessories (and Options)* like *sun roof*, *xenon light* or *rear view camera*. These are boolean attributes since they can either be configured (1) or not (0). For these attributes no price information is given in the dataset.

For an overview of the car configuration attributes, their levels and the first-configuration-prices, see the tables 2 and 3.

Table 2: Overview of all non-binary car attributes. The first column specifies the numeric attribute code and each attribute level is preceded by its numeric code. Only first configuration prices are included.

#	Attribute	No. of levels	Attribute levels (price)
1	Engine	9	1: Engine 1 2: Engine 2 3: Engine 3 ⋮ 8: Engine 8 9: Engine 9 (Price information not provided)
2	Line	4	1: No Line 2: Sports Line 3: Luxury Line 4: Modern Line (Price information not provided)
3	Color	12	1: Alpine White (+0€) 2: Black (+0€) 3: Black Saphire Metallic (+840€) 4: Bluewater Metallic (+840€) 5: Glacier Silver Metallic (+840€) 6: Hematite Grey Metallic (+840€) 7: Crimson Red Metallic (+840€) 8: Mineral White Metallic (+840€) 9: Orion Silver Metallic (+840€) 10: Peacock Blue Metallic (+840€) 11: Sparkling Bronce Metallic (+840€) 12: Deep Sea Blue Metallic (+840€)

⋮ Continued on next page ⋮

Table 2 – Continued from previous page

#	Attribute	No. of levels	Attribute levels (price)
4	Rims	24	1: 16-inch steel basis (+0 €) 2: 16-inch alu basis I (+660 €) 3: 17-inch alu basis I (+1320 €) 4: 18-inch alu basis I (+1980 €) 5: 16-inch alu basis II (+0 €) 6: 17-inch alu basis II (+660 €) 7: 18-inch alu basis II (+1320 €) 8: 17-inch alu basis III (+0 €) 9: 18-inch alu basis III (+660 €) 10: 17-inch alu sport I (+0 €) 11: 18-inch alu sport I (price missing) 12: 17-inch alu sport II (+0 €) 13: 18-inch alu sport II (price missing) 14: 18-inch alu sport III (+0 €) 15: 17-inch alu luxury I (+0 €) 16: 18-inch alu luxury I (+660 €) 17: 17-inch alu luxury II (+0 €) 18: 18-inch alu luxury II (+660 €) 19: 18-inch alu luxury III (+0 €) 20: 17-inch alu modern I (+0 €) 21: 18-inch alu modern I (+660 €) 22: 17-inch alu modern II (+0 €) 23: 18-inch alu modern II (+660 €) 24: 18-inch alu modern III (+0 €)
5	Upholstery	16	1: Fabric Anthracite (+0 €) 2: Leather Dakota Black I (+1750 €) 3: Leather Dakota Veneto Beige I (+1750 €) 4: Fabric Imola Anthracite with Red Contrasting Seam (+0 €) 5: Fabric Imola Anthracite with Grey Contrasting Seam (+0 €) 6: Leather Dakota Black with Red Contrasting Seam (+1750 €) 7: Leather Dakota Everest Grey with Black Contrasting Seam (+1750 €) 8: Leather Dakota Coral Red with Black Contrasting Seam (+1750 €) 9: Fabric-Leather-Combination Anthracite (+0 €) 10: Fabric-Leather-Combination Oyster (+0 €) 11: Leather Dakota Black with Contrasting Seam in Dark Oyster (+1750 €) 12: Leather Dakota Oyster with Contrasting Seam in Dark Oyster (+1750 €) 13: Fabric Salome Saddle Brown / Anthracite (+0 €) 14: Leather Dakota Black II (+1750 €) 15: Leather Dakota Saddle Brown (+1750 €) 16: Leather Dakota Veneto Beige II (+1750 €)

⋮ Continued on next page ⋮

Table 2 – Continued from previous page

#	Attribute	No. of levels	Attribute levels (price)
6	Trims	11	1: Matt Satin Silver (+0 €) 2: Aluminum with Fine Longitudinal Grain and Black Accent Strip (+340 €) 3: Fine-Wood Burr Walnut with Black Accent Strip (+460 €) 4: High Polish Black with Red Accent Strip (+0 €) 5: Aluminum with Fine Longitudinal Grain with Red Accent Strip (+190 €) 6: Aluminum with Fine Longitudinal Grain with Black Accent Strip (+190 €) 7: High Polish Cashmere Silver with Accent Strip in Milky Glass Look (+0 €) 8: Aluminum with Fine Longitudinal Grain with Accent Strip in Milky Glass Look (+190 €) 9: Fine-Wood Fineline Porous Structured with Accent Strip in Milky Glass Look (+310 €) 10: Fine-Wood Burr Walnut with Accent Strip in Chrome (+0 €) 11: Fine-Wood Fineline Anthracite with Intarsia and Accent Strip in Chrome (+100 €)

Table 3: Overview of all binary car attributes. All attributes are preceded by their numeric code.

Attributes
7: Sports package
8: Comfort package
9: Storage package
10: Light package interior
11: Four wheel drive
12: Automatic transmission
13: Cruise control with braking function
14: Cruise control with stop go function
15: Parking assistant
16: Rear view camera
17: Lane change warning
18: Lane departure warning
19: Road sign recognition
20: Head up display
21: Adaptive chassis with lowering
22: Variable sports steering
23: Xenon light
24: Adaptive cornering light
25: Glass sunroof
26: Sun protection blind
27: Sport leather steering wheel
28: Performance leather steering wheel
29: Sports seats for front seats
30: Seat heating for front seats
31: Electric seat adjustment
32: Lumbar support for front seats
33: Climate control
34: Alarm system
35: Arm rest for front seats
36: Comfort access
37: Hitch
38: Navigation system business
39: Hifi system
40: DVD changer
41: Mobile phone prep with bluetooth usb
42: Digital radio

Table 4: A detailed description of the `configuration_type` relation.

Field	**Value range**
type_id	Integer
engine	1,. . . ,9
line	Respective attribute level names from table 2
color	Respective attribute level names from table 2
rims	Respective attribute level names from table 2
upholstery	Respective attribute level names from table 2
trims	Respective attribute level names from table 2
sports_package	Boolean
comfort_package	Boolean
storage_package	Boolean
light_package_interieur	Boolean
four_wheel_drive	Boolean
automatic_transmission	Boolean
cruise_control_with_braking_function	Boolean
cruise_control_with_stop_go_function	Boolean
parking_assistant	Boolean
rear_view_camera	Boolean
lane_change_warning	Boolean
lane_departure_warning	Boolean
road_sign_recognition	Boolean
head_up_display	Boolean
adaptive_chassis_with_lowering	Boolean
variable_sports_steering	Boolean
xenon_light	Boolean
adaptive_cornering_light	Boolean
glass_sunroof	Boolean
sun_protection_blind	Boolean
sport_leather_steering_wheel	Boolean
performance_leather_steering_wheel	Boolean
sports_seats_for_front_seats	Boolean
seat_heating_for_front_seats	Boolean
electric_seat_adjustment	Boolean
lumbar_support_for_front_seats	Boolean
climate_control	Boolean
alarm_system	Boolean
arm_rest_for_front_seats	Boolean
comfort_access	Boolean
hitch	Boolean
navigation_system_business	Boolean
hifi_system	Boolean
dvd_changer	Boolean
mobile_phone_prep_with_bluetooth_usb	Boolean
digital_radio	Boolean

References

[cyLEDGE Media, 2018] cyLEDGE Media (2018). Configurator database. `www.configurator-database.com`. Accessed: 2018-03-24.

[Fuhrmann et al., 2016] Fuhrmann, T., Schweizer, M., Geyer-Schulz, A., and Kurz, P. (2016). Mining consumer-generated product-configuration data. (Manuscript submitted for publication.).

[Fuhrmann et al., 2017] Fuhrmann, T., Schweizer, M., Geyer-Schulz, A., and Kurz, P. (2017). On estimating pricing models from end-consumer internet car-configuration data. *WIAS Reports*, 29(1):55 – 69.

[Mittal and Frayman, 1989] Mittal, S. and Frayman, F. (1989). Towards a Generic Model of Configuraton Tasks. In *IJCAI*, volume 89, pages 1395–1401.

[Rüde et al., 2016] Rüde, R., Fuhrmann, T., Neureuther, T., Lausen, L., Schweizer, M., and Geyer-Schulz, A. (2016). Mining consumer-generated product-configuration data. Presented at Tagung der Deutschen Arbeitsgemeinschaft Statistik (DAGStat), Göttingen.

[Schweizer, 2018] Schweizer, M. (2018). Analyzing economic choice with consumer-generated product-configuration data.