IDA Case Study 71

Your group works for an independent risk assessment company. A famous car magazine wants your company to evaluate failures of cars manufactured by the company OEM2 from the latest two production years for an article. The magazine provides you with data from different sources of each level of the supply chain for the analysis. Because the content of the article is not clear now, they want you to provide them an application containing different approaches for failure analysis.

Information about tasks and data sets:

All information, including when a vehicle was produced in which factory and if a vehicle is defective, can be found in the group's manufacturing data.

A vehicle is considered as defective if a part, a component or the whole vehicle is marked as defective. The same logic applies for components containing defective parts.

You get access to the group's own database and have to decide which data you need for your analysis. In addition, the Federal Motor Transport Authority (KBA) gave you registration data and geodata which are also stored in the database. Here are the categories of datasets:

- Part = Einzelteil
- Component = Komponente
- Vehicle = Fahrzeug
- Geodata = Geodata
- Registrations = Zulassungen
- Logistics delays = Lieferverzüge

It is recommended to have a look at the structure of the single tables before starting the analysis.

For all instances of the supply chain, i.e. for parts, components and vehicles, production dates with information about ID number, manufacturer, production factory, production date and defect entries are available.

The ID numbers consist of part designation, manufacturer, factory and consecutive number. Example: 1-201-2011-3, Part T1, produced by manufacturer "201" in factory "2011", 3rd part from this series.

For components and vehicles, item lists are available, which are declared by the naming convention element_name_shortcut. The lists contain information about all installed parts resp. components.

Hints for the approach to your case study:

The basic target is the development of an application to analyze your problem. It should be possible to manipulate certain settings interactively, that are automatically considered while analyzing. To evaluate the results, it is important to document your approach with an R Markdown file. Follow the steps below:

1. Import relevant datasets from the provided tubcloud folder. Start this task by listing all documents which are required.

2. Prepare the data according to the principles of *tidy data* and put them together in a single dataset.

3. Develop a Shiny-App which fulfills the following criteria:
    - All used packages must be loadable by the current R-version.
    - The Application must be executable from the submission folder without any additional accommodations. It is advisable to test for this behavior before submitting.
    - The application should only refer to the dataset you created in step two.
    - The layout of the application should be in accordance with the target group and the corporate color, which is grass green. Furthermore, a logo should be integrated in the layout. You can either design your own logo or you can use the logo of the department for quality science (Fachgebiet für Qualitätswissenschaften). In addition, the font must be changed in compliance with your preferences. A change of font is mandatory.

4. Develop a Shiny app and visualize the following:
    a. All failed cars from the latest two years of production of the available data in a chart combining two representations in form of bars and line plots. The user should be able to switch between different representations of the two years (per month, quarter, and years). The bars and lines must contain interactive elements showing the number failures. Integrate an additional filter option for the type of cars.
    b. On the same page the customer wants you to provide an additional chart representing the average mileage per vehicle type before failure connected to the first chart. This chart must change according to the time setting of the first chart.
    c. Add an additional page containing a map which contains the location of factories of the company OEM2 and its suppliers. Add pop ups showing the total number of produced and failed cars, components, and parts. Highlight the factories which have the highest failure rates.
    d. Your dataset on which your analysis is based so you can prove your visualization. Remember: Only show necessary attributes. Therefore, add a table to your application, containing only necessary content.

5. Document the results of your analysis in an R Markdown file. Describe your data analysis process step by step and discuss the results by referencing convincing graphics from your developed app. The description of your approach is necessary for the traceability of your solution steps. If there are problems with your R code, the documentation can be used to acknowledge a correct approach. Therefore, comment your code carefully and keep the tidyverse style guide in mind. Create a clickable index which should be always displayed in the html-file. The index should at least contain the following sections:
    - Data import
    - Data preperation
    - Creation of final detaset
    - Evaluation
    - Results

6. The submission folder has to contain the following files. XX is to be replaced with the group number:

- Final_Data_Group_XX.Rdata
- General_Tasks_Group_XX.rmd
- General_Tasks_Group_XX.html
- Case_Study_Group_XX.rmd
- Case_Study_Group_XX.html
- Case_study_App_Group_XX.r

General tasks

In addition to the case study, there are further general tasks to do. Remember the documentation of your code, so that your solutions can be understood and if necessary, points for incomplete solutions can be given.

1. Logistics play a more and more important role in the product development of the automobile industry. Parts produced by the supplier must first be delivered to the OEM before they can be installed. What seems logical at first sight should be analyzed in more detailed way for a professional application.
   Therefore, create a distribution for the logistics delay of component „K7". Use the production date (“Produktionsdatum”) from the data set "Komponente_K7.csv" and the receiving date of incoming goods (“Wareneingang”) from "Logistics_delay_K7.csv". You can assume that the produced goods are issued one day after production date.
   For the model design in R, create a new data set "Logistics delay" that contains the required information from both data sets.
   Answer the following questions:
   - How is the logistics delay distributed? Proof your selection by statistical tests and briefly describe your approach.
   - What is the minimum/maximum time between delivering and receiving goods?
   - Determine the mean of the logistics delay.
   - Visualize the distribution in an appropriate way by displaying a histogram and the density function using the package plotly.

2. Why does it make sense to store the available data in separate files instead of saving everything in a huge table? Name at least four benefits. The available tables represent a typical data base structure. How is it called?

3. How many of the parts T4 ended up in vehicles registered in the city of Dortmund?

4. Which data types do the attributes of the registration table "Zulassungen_aller_Fahrzeuge" have? Put your answers into a table which is integrated into your Markdown document.

5. You want to publish your application. Why does it make sense to store the data sets in a database on a server? Why is it not recommended to store the data sets on your personal computer? Name at least three points per question.

6. On 11 August 2019 there was an accident involving a stolen car produced by your company. The driver left the scene without a trace. The license plate of the car, which caused the accident, was faked and the Vehicle Identification Number (VIN) was removed. Since you work for the Federal Motor Transport Authority, the police asks for your help to find out where the vehicle with the engine code "K1DI2-103-1031-21" (corresponds to the engine ID number) was registered.