ISyE 6740 – Fall 2023

Project Proposal

# TEAM **MEMBERS**

**Hassan Abdrabo (GTID: 903649924)** is a seasoned data analyst at New York University Abu Dhabi with a bachelor's degree in information technology.

**Duc Nguyen Hong (GTID: 903943817)** holds a bachelor's degree in Finance & Banking and has established himself as a senior data analyst at an outsourcing company based in Vietnam.

**Fuad Jannatov (GTID: 903847334)** is an engineer at Baker Hughes in Oil and Gas industry with a bachelor's degree in petroleum engineering.

# PROJECT TITLE:

*Car price estimation in German and Czech Republic.*

# PROBLEM STATEMENT

**Background Info:** The car market in Europe is a dynamic and multifaceted sector that plays a pivotal role in the region's economy and daily life. Known for its rich history of automotive innovation, Europe has established itself as a global hub for both car manufacturing and consumption. This project will focus on the German and Czech market, highlighting their key features and price trends.

**Problem Statement:** Demand for automobiles had been increasing for years. In 2017, the European passenger car market experienced a 3% growth, reaching almost 15.5 million in sales. This marks the fourth consecutive year of expansion in Europe and represents the highest sales volume since 2007. Among the prominent markets, Italy, Spain, and France stood out with growth rates of 7.9%, 7.7%, and 4.7%, respectively, surpassing the overall market performance, while Germany also experienced a solid growth rate of 2.7%[[1]](#footnote-2).

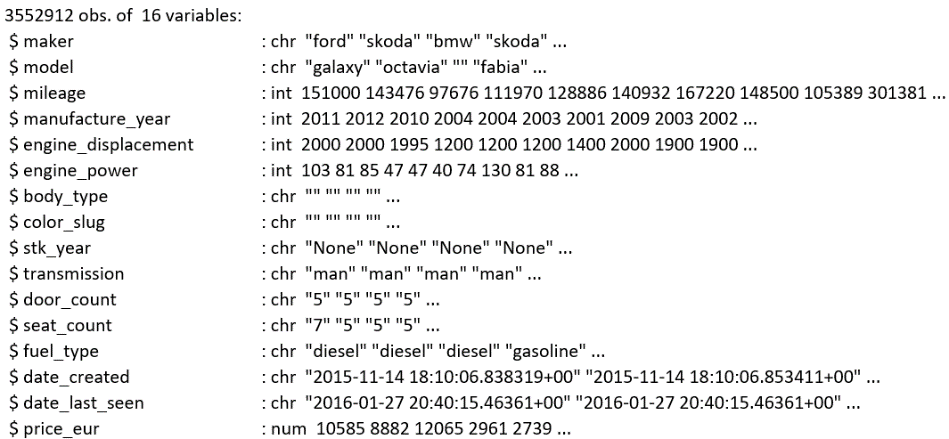
Estimating prices is a crucial task for both buyers and sellers to facilitate informed decisions in terms of trading opportunities. Furthermore, meaningful analyses contribute to market transparency, fair pricing practices, and a more efficient and equitable automotive marketplace.

**Primary Research Problem:** Car price estimation based on data scraped from advertisements.

**Complementary Questions:**

* Which factors determine the price of a car?
* With what accuracy can the price be predicted?
* Can a model trained on all cars be used to accurately predict prices of models with only a few samples?

# DATA SOURCE

The dataset contains information scraped off various websites in Czech Republic and Germany from 2015 to 2017 by Kaggle user [Miroslav Zoricak.](https://www.kaggle.com/mirosval)

There are 3.5 million raw datapoints in the dataset, each data point representing a scraped from website ad of a car, having 15 independent and 1 dependent (price\_eur) features. The data types are mixed, containing strings, integers, and datetime types.

# METHODOLOGY:

**Exploratory Data Analysis:** Our goal is to use the present information about a car to predict its value. The details available for each car vary, as the dataset has a high number of null values in each column. Hence, our models must be flexible, and able to fit the data well, to provide a realistic estimate for the car's value based on the available information.

After an initial analysis, we have also observed that our dataset has many extreme values and mismatches in specifications for cars of the same model. This could be an issue in the scrapping of data or a normal occurrence due to changes in the model specification. We plan to invest a significant amount of time exploring the dataset and planning for the challenges we might face in the modeling process.

**Preprocessing:** Before this dataset is ready for analysis, appropriate methodologies for data cleaning and wrangling must be implemented. The main challenges present within this dataset are null and extreme values. The dataset will be split among the team members where each member will be responsible for developing rational methodologies for cleaning the data. The strategies implemented for handling such issues will vary based on the issue and the data. Such strategies include, but are not limited to, imputing missing values, building separate models to handle such values, and dropping observations with a high number of missing or extreme values.

To better understand our data, we also plan to implement multiple unsupervised learning algorithms for identifying patterns in the data like K-Means clustering and identifying outliers like PCA.

**Modeling:** To model car prices, we plan on implementing testing different algorithms. We will start with Multiple Linear Regression and apply it on the complete dataset after cleaning. We also plan on testing other regression approaches including transforming the response variables and the predictors and different types of Generalized Linear Models that assume different distributions for the data like the Gamma distribution. Additionally, we plan on trying other modeling techniques including K-Nearest Neighbors, Decision Trees, Random Forests, Gradient Boosting, and Neural Networks.

**Anticipated Conclusions & Hypothesis:** We believe that make, model, mileage, manufacture year, transmission and fuel type affect the price of the car. We will test this hypothesis in our analysis.

# EVALUATION & CONCLUSION

**Evaluation:** We will split our data into 3 sets: one for learning, one for validation and model selection, and the final one for testing and performance evaluations. To evaluate models' performance, we are going to implement different evaluations methods like residual analysis and report various metrics like MSE, MAE, RMSE, and adjusted-R2.

**Business decisions impacted and Benefits:** Understanding the most important factors that determine a car’s price can be used to develop innovative applications that would suggest a fair price for the car to the seller and the buyer. We plan on implementing a web-based application that will provide recommendations for car prices in the future.

# PROJECT TIMELINE/PLANNING Project Timeline/Mention key dates we hope to achieve certain milestones by:

|  |  |  |
| --- | --- | --- |
| **Phase** | **Description** | **Timeline** |
| 1 | Team Formation | 09/19 |
| 2 | Finding a Project | 09/19 |
| Business Understanding & Feasibility | 09/26 |
| Data Preparation & Simple EDA | 09/26 |
| Project Proposal Submission | 10/17 |
| 3 | Deep Exploratory Data Analysis | 10/24 |
| Modeling & Validation | 10/31 |
| Final Report Write-up | 11/07 |
| Final Report Submission | 11/28 |

1. https://www.goodcarbadcar.net/2017-europe-sales-figures-by-brand/ [↑](#footnote-ref-2)