**ISYE 6414 Project Abstract - Team 18**

**The student names in the team**

* Beenapreet Kaur Singh
* Rachel Ramsay
* Eugene Kang

**A brief description of the data and the source where you acquired the data**

Our project will use Kaggle’s ‘Used Car Database’ scraped from Ebay-Kleinanzeigen (German second-hand website). This dataset contains about 370,000 observations of used car advertisement posted on the website from March 2015 to April 2016. Specifically, dataset includes the attributes such as prices, vehicle types, models, brands, fuel types, etc. The column lastSeen and dataCreated will be utilized for estimating how long a car advertisement will be online before it is removed from the page. One of the concerns related to this dataset is that it has a lot of missing values. To make our analysis concrete, we need data cleaning so that we can draw a reliable conclusion. Our group would like to further subset this data into a smaller, tidy dataset on which we will perform analysis.

**A brief description of the questions that you would like to answer**

Since we all three live in Atlanta and rely on cars to get many places, we are interested in what we can find out about purchasing, repairing, and selling used cars from this data. A few questions we want to ask are listed below:

1. If I get a notification that a car I like is on eBay, is it likely to stay on the site for a day, or should I go ahead and put a bid in immediately before the post is taken down?
2. I want to sell a car that has damage. Should I repair the damage first? (Given cars of the same year and model, how much more do repaired calls sell for than damaged cars?)
3. I am a loyal Volkswagen customer, and want to buy a VW with the highest resell value. Which model should I buy?
4. If I have a strict budget of $10,000, how old are the cars in my budget likely to be?
5. If I only want to purchase a car that is less than 3 years old, how much should I budget for that car?
6. ADDED LATER\*\* check in on all caps, exclamation points, length of string (number of \_ or chars)

**A brief description of the steps we will take in performing the analysis:**

**Data Cleaning**

Remove rows with NA/ missing values or impute data for those observations

Remove columns with more than 30% of data with missing values

Convert German characters to english

Parse Date values in proper format for analysis

Convert categorical variables to factors in R

Decide if and how to subset the data to reduce size (perhaps on car model)

**EDA- Exploratory Data Analysis**

1. Find how price of ad varies with different brands and model of car for different levels of damage -(high, medium , low - self defined dummy variable) (answers qn 2)
2. Find correlation between price on ad and the different variables - (Hypothesis: Price on ad is highly correlated with whether of not the car was damaged or Price is highly correlated with the engine power of the car)
3. Check distribution of cars for different budgets - find out how price varies with age of car - simple linear regression (answers qn 4)
4. Classification of used Car brands by the time its ad stays online For Example - 1-10 days, 11-20 days - This tells us for which car we need to put a bid immediately (answers qn 1)
5. More exploratory analysis as we see fit.

**Main Analysis**

**1) To decide which car to buy given certain specifications by user (questions 3-5)**

Fit a linear regression model with response variable as price. This model will also help understand how much one can expect to pay for a particular brand or model of car given its specifications.

1. Check the distribution of different features of used car with the price on ad
2. Check if all assumptions of linear regression model hold - if not transform variables to get the best fit.
3. Use correlation chart to choose the best predictor variables. If 2 important variables are highly correlated use either stepwise regression/ lasso or elastic net to choose the best variables - compare models with best adj R square.

**2) To decide whether repairing a damaged car is worth the increase in resell price (question 2)**

Qn - how much less does a model x car sell for if it is damaged, vs all things constant and not damaged? This would help a seller decide the value of performing a repair on the car before selling it. We will perform Multiple Linear Regression for this task.

**Other steps we will take within our analysis:**

1. Fit a range of models to explore question and perform Goodness of Fit analysis to confirm that our models fit the dataset well
2. Perform our model’s analysis on a test set of data from the same dataset

**Tasks:**

1. Data Cleaning - Eugene
2. EDA - Beena
3. Main Analysis 1 - all
4. Main Analysis 2 - all
5. Additional steps - all
6. Project Report writing - Rachel