# Phase 1: Proposal: Dataset and Topic Selection

# Team Members

* Emma Doyle, Reid Hoffmeier, Matthew Lane, Mikhail Mikhaylov, Brandon Strong

# Dataset and Topic

* Dataset: Kraggle Used Cars Market in Belarus
* Topic: This dataset was collected on December 2nd 2019 from one of the most popular online catalogs in Belarus. We will use this dataset to find trends and patterns that relate to geographic, economic, and regional influences in Belarus.

# Preparing and cleaning the dataset

* Using R, Mikhail’s understanding of the language, and Google Translate
  + Translate, recode, and mutate the data into English.
* Check for null values and outliers.
* Present data as graphical representations to answer our questions.

# Questions and Goals

* What is the distribution of manufacturers for each region and whether manufacturer has a significant impact on the asking price of a vehicle?
  + **Solved by: Pie graph**
* What is the most popular model by region and whether we can conclude that the popularity of a model has a direct impact on the price of a vehicle?
  + **Solved by: Bar graph**
* What is the average age of each vehicle manufacturer and whether manufacturer has an impact on the impact of production year on the selling price?
  + **Solved by: Group by, Summarize, Bar graph**
* What is the average asking price for each region and what impact does a region have on price, does body type make a difference on the asking price for a region?
  + **Solved by:** **Scatter Plot**
* What is the relationship between odometer and price and whether sellers are more likely to accept exchanges as the odometer value increases (Fill in with whether its exchangeable)?
  + **Solved by: Scatter Plot**
* What is the relationship between engine type and engine capacity and what is the impact of this on the selling price?
  + **Solved by: Scatter Plot**
* Manufacturer Origin Distribution
  + **Solved by: Pie graph**
* Distribution of Engine type for each region and by car type?
  + **Solved by: Mosaic Plot**
* Distribution of colors to body type.
  + **Solved by: Bar Graph**
* Which variable has the largest impact on the selling price of a vehicle what impacts could explain the outliers for this correlation?
  + **Solved by: Inspection of all previous graphs**

# Collaboration Method

* **Discord** used for group chat and virtual meetings as needed.
  + Group response typically received within 24-48 hours of questions asked.
* **Github** used for code/document sharing.
* **Google Docs** used for simultaneous document editing and idea pooling.