**SECRET LINK** - <https://leoportfolio.acu.edu.au/view/view.php?t=MHLQbFvN87deO9z6ZXmS>

Bottom of Form

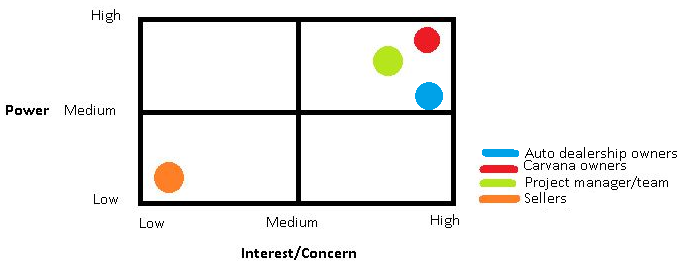
Bottom of Form

**DATA201 Reflection**

**Introduction**

A “Kick” in the auto auction industry occurs when cars are sold to auctioneers and are then later un-sellable, making the auctioneers lose money. The supplied Excel sheet is aimed at displaying the pre-processed and analysed data consolidated into relevant information. Trends and correlations will be identified between kicked and non-kicked cars. This document is aimed at providing a description of techniques used to engage stakeholders so that relevant data can be extracted. It is also aimed at describing data pre-processing and data analyses carried out to arrive at conclusions stipulated by the processed spreadsheet information. Spreadsheet data will be used to attempt to create a tool for identifying if cars will be kicked or not.

The primary external stakeholders are: the owners of the auction company Carvana (Carvana n.d.) and the auto dealerships purchasing the potentially kicked cars. Other internal primary stakeholders are the data analyst manager and the data analysis teams. Any sellers are secondary stakeholders. It is important to note that only some stakeholders have power and interest in the project:



The project manager/team interviewed the owners to obtain a goal for what was expected of the project. Then the project manager and team brainstormed to theorise what information would be of use to them.

**Questions:**

**What do you want to see change in the system?**

**Do you have any theories of what car may be kick cars or not?**

**What is the usual process of the car purchase process?**

**Data pre-processing**

Data pre-processing is a set of steps carried out to help make data more user- friendly format. The steps are data consolidation, data cleaning, data transformation, and data reduction.

**Data consolidation**

There are many different datasets within the raw data spreadsheet. The job of a data analysis team is to make data more user friendly. Consolidation allows the data to become more user friendly for the data analysts so that they can do so for others. Firstly, data that has no possible correlation with whether the car is kicked or not. The data NOT collected from the main spreadsheet is:

* **RefId** – This data is for organising the complete range of data, which will be getting cleaned, transformed and reduced anyway. If included, it will be unusable by the end of the process.
* **Auction** – The auction provider at which the vehicle was purchased cannot legitimately be concerned with the sellable condition of the car.
* **SubModel** – This goes into too much detail, and will possibly make false correlations as a result.
* **Trim** – Trim is the term used to describe cosmetic features of a car. Because it is cosmetic, there is no condition the trim can be in to warrant the car as unsellable.
* **Color** – Excluded for same reasons as **Trim**
* **WheelTypeID/Wheel Type** – Excluded for the same reasons as **Color** and **Trim**.
* **MMRCurrentRetailAveragePrice/** **MMRCurrentRetailCleanPrice** – Because the analysis is only concerned with the auction part of the business process, and not the retail price of the car, the acquisition price of the car in the retail market is not relevant.
* **Kickdate** – The date that the vehicle was kicked back to the auction does not provide any information about *why* it was kicked.
* **BYRNO** – A placeholder to assign a unique number to customers who were kicked. For the same reasons as **RefId**, it cannot be used.
* **IsOnlineSale** – All of the values in this column are “0”, so are effectively valueless
* **VNZIP1/VNST** – location is irrelevant to conditions that could affect whether the car is a kick or not.

**Data cleaning**

The supplied raw data has various duplicates, null and void values and other unusable data. Where more than 10% of values in a single column return “NULL” or “VOID” values, the data of the entire column will not be used as the size of the missing sample may make a significant outcome to the analysis results. Where less than 10% of values are “NULL” or “VOID”, the row containing these values will be eliminated. If a duplicate is found, the duplicate with information deemed more suitable for the analysis at the project manager’s discretion will be chosen.

* **VehYear/VehicleAge** – These columns are essentially duplicates. In this case, I will remove the **VehicleAge** column, as the data will only be correct for one year, but **VehYear** identifies the year of manufacture, so remains stationary.
* **Nationality** – The nationality of the car does not directly pertain to whether it will be a kick or not. This is because the **Make** column describes sub-categories of **Nationality** with greater detail which allows for more highly refined information in the finished product of this analysis.
* **TopThreeAmericanName** – Cleaned for the same reason as **Nationality**. The data can be described in greater detail by the **Make** column.
* **PRIMEUNIT** – This column was found to have 95.3% of its data as “NULL”, making the rest of its data unusable.
* **AUCGUART** – Unusable for the same reasons as **PRIMEUNIT**.
* **Transmission** – It was calculated with percentage techniques that automatic transmissions accounted for 96% of entries. Because kicked cars only make up 12.3%, there can be no meaningful correlations.
* **Size** – The **Make** and **Model** The size of the vehicle is not relevant to the condition of the car.

**MMRCurrentAuctionAveragePrice/MMRCurrentAuctionCleanPrice** – Several “NULL” iterations were found, but made up less than 10% of the total column data, so the rows containing the “NULL” data were removed. Also identified in **MMRCurrentAuctionAveragePrice** and **MMRCurrentAuctionCleanPrice** are 139 iterations with “0” as their cell value. This data will be eliminated.

The remaining columns are:

**IsBadBuy, PurchDate, VehYear, Make, Model, VehOdo, MMRCurrentAuctionAveragePrice, MMRCurrentAuctionCleanPrice, VehBCost**

Now the data will be categorised into two main sheets – **IsBadBuy0/1**. These two datasets contain the now-consolidated and cleaned data. This is achieved by putting kicks and non-kicks onto separate spreadsheets. After this, **IsBadBuy0/1** columns will be eliminated

Now outliers will be identified. First of all, averages will be determined from each column of **IsBadBuy0/1.** Then, when a total of each column occurrence has been established, averages of occurrence are calculated. This processing reveals that some values in the **Make** columns are outliers because they make up less than 10% of the total. The makes of car that make up 10% or more of the total are:

* **Chrysler** – 12.2%
* **Ford** – 15.5%
* **Dodge** – 17.7&
* **Chevrolet** – 23.7%
* **Total** – 75.4% of total car makes are attributed to these four makes. All other **Make**s are removed.

**Data transformation**

The data has now been processed to the point where it can be used as meaningful information.

Histograms will be used to visually interpret trends with **VehYear** between **IsBadBuy0** and **1**. **Make** of **IsBadBuy0** and **1** will be processed in the same way. Both columns have the percentages calculated from the **VehYear** and **Make** cell values so that they can be figuratively compared as well as literally.

**MMRCurrentAuctionAveragePrice** and **MMRCurrentAuctionCleanPrice** will be interpreted using a box and whisker diagram for both **IsBadBuy0** and **1**. This allows these two columns to be easily visually interpreted but forgoes much of the consolidation and de-duplication process. **VehOdo** will be transformed via identical method.

**VehBCost** can be run against **MMRCurrentAuctionAveragePrice** and **MMRCurrentAuctionCleanPrice** to analyse whether cars in **IsBadBuy0/1** are in average or clean condition.

**Data reduction**

Data reduction is the point where the meaningful information becomes user friendly. This is done with a series of volume and dimension reductions, and data normalisation.

Up until this point, The data I have been working with has been very messy. So first of all, I will be consolidating all relevant findings in a sheet named “Stakeholder Product”, and the column names will be renamed to more easy-to-understand labels. The rest of the data will remain in the sheet for transparency and stakeholder perusal.

**Conclusion**

Unfortunately, a tool to predict cars that will be kicks was unobtainable. To an extent, predictive analysis has been made, and from this the following conclusions can be reached:

**Kicked cars:**

* Are most likely to be a Ford.
* Are most likely manufactured in 2005.
* See roughly a 1:3 ratio for average condition and clean condition.
* Have a lower average and clean condition car price.
* Tend to have lower odometer readings than non-kicked cars.

**Non-kicked cars:**

* Are most likely to be Chevrolets.
* Are most likely manufactured in 2006.
* See roughly a 1:1 ratio for average and clean condition.
* Have a higher average and clean condition car price.

**References**

Carvana n.d, *About Us*, Viewed 2 September,

<<https://www.carvana.com/about-us>>

**Bibliography**

Sharda, R Delen, D Turban, E 2018, *Business Intelligence, Analytics, And Data Science*, 4th edn, Pearson, New York,