Kyndall Butler

Final Project Written Report Score: \_\_\_\_\_\_\_\_ of 100

Computer STAT 3010 Section 51

# Introduction

Buying used cars has become increasingly more common as it is typically the most economically plausible route for buying a car. The objective of this analysis was to determine if there is a relationship between the distance a car has traveled and the car’s selling price and to determine if there is a relationship between the year the car was released and the car’s selling price. The dataset consists of 2020 used car sales in India from a used car website. The collected data will contribute to how differently cars are valued internationally. I am in the process of finding and purchasing a used car; it is interesting to see how this process would differ across the globe.

### Statistical Question 1: Bivariate Analysis of Two Qualitative Variables

Are certain fuel types more popular in earlier years than later years?

### Statistical Question 2: Bivariate Analysis of a Quantitative Variable and a Qualitative Variable

Does the number of kilometers driven vary by fuel type?

### Statistical Question 3: Bivariate Analysis of Two Continuous Variables

Is the car’s selling price independent of the number of kilometers driven?

# Section 1: Data set overview and preparation

## 1.1: Data set overview

Filename: Cars Extension: [Cars.csv](https://www.kaggle.com/nehalbirla/vehicle-dataset-from-cardekho/download)

## 1.2: Data set summary

This data set was collected in India and was sourced from a website that allows customers to buy and sell used cars in India. The dataset was collected to analyze trends in car sales in India. The original dataset has 302 records and nine variables.

## 1.3: Data preparation methods

I created a subset (or new set) of data that included 6 variables and 288 records. The subset excludes classic cars that are extreme outliers because they do not match the tier of the other cars. Within the subset I created an ordinal variable– yeargroup, from the discrete variable ‘year’. The ordinal variable had four groups with group one being the oldest era and group four being the newest era. The subset included: the car’s name, the number of kilometers driven, the fuel type, the car’s selling price, yeargroup, and the car’s year. I chose these variables because I was interested in exploring two factors that influence the car’s selling price as well as the factors that influence the two factors of interest.

## 1.4: Table of variable definitions

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variables | Definition | Values | Type | Measurement Unit |
| Car Name | The model of the car. |  | Qualitative  (nominal) |  |
| Yeargroup | The era the car was bought and released. | 2003- 2006= 1  2007-2010= 2  2011- 2014= 3  2015- 2018= 4 | Qualitative  (ordinal) | Years |
| Year | The year the car was bought and released. | 2003 => year => 2018 | Quantitative  (discrete) | Years |
| Fuel Type | The type of fuel the car requires. | CNG, Petrol, or Diesel | Qualitative  (multinominal) |  |
| Kilometers Driven | The number of kilometers the car has been driven. | 1. => x => 500,000 | Quantitative  (continuous) | kilometers (km) |
| Selling Price | The amount of dollars the car was sold for. | * 1. => x => 0.35   (values were divided by 1,000) | Quantitative  (continuous) | dollars |

# Section 2: Univariate descriptive statistics and visualizations

## 2.1: Ordinal variable

Table 1. *Table of frequency, relative frequency, cumulative frequency, and cumulative percent.*

| **yeargroup** | **Frequency** | **Relative Frequency** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **4** | 138 | 0.4792 | 138 | 47.92 |
| **3** | 110 | 0.3819 | 248 | 86.11 |
| **2** | 29 | 0.1007 | 277 | 96.18 |
| **1** | 11 | 0.0382 | 288 | 100.00 |
| **Total** | 288 | 1 |  |  |

Figure 1. *Bar chart of year groups.* Figure 2*. Ordered bar chart of year groups.*

Interpretations

The oldest era of cars (group one) was sold the least frequently which is to be expected considering the data was collected in November of 2020, so the demand for newer cars will be higher. Table 1, figure 1, and figure 2 support this claim as there is a noticeable trend– the newer the era, the higher the frequency. Models released prior to 2011 were sold significantly less the cars released during and after 2011.

| **Table of yeargroup by Fuel\_Type** | | | | |
| --- | --- | --- | --- | --- |
| **yeargroup** | **Fuel\_Type** | | | |
| **CNG** | **Diesel** | **Petrol** | **Total** |
| **1** | 0  0.0000  0.0000  0.0000 | 1  0.0035  0.0909  0.0204 | 10  0.0347  0.9091  0.0422 | 11  0.0382 |
| **2** | 0  0.0000  0.0000  0.0000 | 3  0.0104  0.1034  0.0612 | 26  0.0903  0.8966  0.1097 | 29  0.1007 |
| **3** | 1  0.0035  0.0091  0.5000 | 24  0.0833  0.2182  0.4898 | 85  0.2951  0.7727  0.3586 | 110  0.3819 |
| **4** | 1  0.0035  0.0072  0.5000 | 21  0.0729  0.1522  0.4286 | 116  0.4028  0.8406  0.4895 | 138  0.4792 |
| **Total** | 2  0.0069 | 49  0.1701 | 237  0.8229 | 288  100.00 |

# Section 3: Multivariate descriptive statistics and visualizations

## 3.1: Bivariate Analysis of Two Qualitative Variables

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedTable 2. *Contingency table of frequency, relative frequency, row percent, and column percent of year groups.*

Figure 4. *100% stacked bar chart of fuel type by year group.*

Figure 3. *Side-by-side bar chart of fuel type by year group.*

Interpretations

Petrol fuel gets increasingly more popular as years pass as does diesel fuel. Because petrol fuel was the most frequently recorded, there is visibly a significant difference in popularity when comparing petrol to other fuels. However, the data does not support the claim that a certain fuel type is more popular during earlier years than later years or vice versa because it is important to recall the frequency of cars recorded in the later years (group 3 and 4) is much higher than earlier years (group 1 and 2). The fuel type will increase in later years simply because cars from later years were more dominant in the study; unfortunately, I did not have this revelation before performing the analysis.

## 3.2: Bivariate Analysis of a Quantitative Variable and a Qualitative Variable

Table 3. *Table of numerical summaries for each fuel type.*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Fuel\_Type** | **N** | **Mean** | **Std Dev** | **Minimum** | **Median** | **Maximum** | **N Miss** | **Lower Quartile** | **Upper Quartile** |
| CNG | 2 | 42749 | 10252 | 35500 | 42749 | 49998 | 0 | 35500 | 49998 |
| Diesel | 49 | 50984 | 28587 | 2071 | 45000 | 197176 | 0 | 40000 | 59000 |
| Petrol | 237 | 33729 | 40420 | 500 | 26000 | 500000 | 0 | 14000 | 44542 |

Chart, box and whisker chart

Description automatically generated

Figure 5. *Boxplot of Kilometers Driven by Fuel Type.*

Interpretations

With 95% confidence the average amount of kilometers driven for cars that require CNG fuel is between 28,500 and 56,900 km, based on only 2 observations. With 95% confidence average amount of kilometers driven for cars that require Diesel fuel is between 43,000 and 59,000 km, based on 49 records. With 95% confidence the average amount of kilometers driven for cars that require Petrol fuel is between 28,600 and 38,900, based on 237 observations.

Cars that require petrol fuel were typically driven less, but this could be a result of more observations (a larger pool of data); this fuel type was recorded most frequently and will have a larger range resulting in a smaller average. This large degree of variability is supported by the large standard deviation (SD= 40420 km) and the outliers in the side-by-side boxplot. The fuel type that had the car that was driven the least in the entire subset of data was the Petrol fuel type with the minimum being 500 kilometers.

## Chart, scatter chart Description automatically generated3.3: Bivariate Analysis of Two Continuous Variables

Figure 6. *Scatterplot of kilometers driven v selling price from simple random sample (n=29)*

*data.*

| **Simple Statistics** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **N** | **Mean** | **Std Dev** | **Sum** | **Minimum** | **Maximum** |
| **Kms\_Driven** | 29 | 37511 | 38728 | 1087805 | 1900 | 213000 |
| **Selling\_Price** | 29 | 3.05241 | 2.85825 | 88.52000 | 0.12000 | 9.25000 |

Table 4. *Table of numerical summaries for continuous variables in the simple random sample (n=29).*

Interpretations

The correlation (r= –0.08279) indicates a very weak negative relationship between the number of kilometers driven and the selling price indicating an inverse relationship. The slope (m= –6.11e-6) also indicates an inverse relationship between the two variables, as kilometers increase, selling price is expected to decrease. When the car has been driven 2,000 kilometers, the price is expected to be 3.29k dollars; we can reference a car that has been driven 2,135 kilometers, it was sold for 2.85k. This proves the slope is not an extremely accurate indicator of the selling price, but it is arguably decent. Selling price is not completely independent of kilometers driven, but there is no significant correlation between the two. As demonstrated in figure 6, selling price is not proportional to kilometers driven, it does not increase as kilometers driven increases or vice versa.

# Section 4: Conclusion

As someone in the process of finding and purchasing a car, the data has not been extremely useful as none of the relationships were significant and the visualizations were misleading. After analysis, I think the value of a car is impacted by the number of kilometers driven more than the car’s ‘age’. The amount of money a car can sell for is influenced by a superfluity of factors that were not explored here like trims, customer location, customer’s age, car color, etc. Despite the cars being very similar in features, MSRP, and tier, I did not consider how factors like dependability, brand loyalty, and miles per gallon would influence the purchase trend. For this analysis to be more valuable I will need to include a more diverse set of data (for example have more cars with CNG fuel type recorded) and more variables like the seller type, the color, the miles per gallon, and the customer demographic. This analysis could be used to prove how inconsistent trends can be, but I do not think it will be valuable for determining any influential factors for a car’s selling price.

# APPENDIX

## Annotated SAS Code

# Section 1

SAS code for data import, data preparation, and created variables.

/\*Create a library connection.\*/

libname fpcars "/home/u58487288/Cars SAS Final Project";

/\*Import the data file and store it in the library.\*/

proc import datafile= '/home/u58487288/Cars SAS Final Project/cardata.csv'

out=fpcars.cardata replace;

/\*Copy the data set.\*/

data fpcars.projectdata;

set fpcars.cardata;

\*Subset the data\*;

data fpcars.fpdata (DROP= transmission present\_price owner seller\_type VAR10 VAR11 VAR12);

set fpcars.projectdata;

IF selling\_price < 15 then output;

IF selling\_price > 15 then delete;

\*create the ordinal variable\*;

data fpcars.projectdata;

set fpcars.cardata;

IF year => 2003 and year <= 2006 then yeargroup= '1';

else if year => 2007 and year <= 2010 then yeargroup= '2';

else if year => 2011 and year <= 2014 then yeargroup= '3';

else if year => 2015 and year <= 2018 then yeargroup= '4';

# Section 2

SAS Code for numerical summaries and visualizations for this section.

\*frequency tables for yeargroup\*;

proc freq data=fpcars.fpdata order=freq;

table yeargroup;

\*scatterplot for yeargroup\*;

proc sgplot data=fpcars.fpdata;

title 'Bar Chart of YearGroup';

hbar yeargroup/ stat=percent;

xaxis label= 'Year Groups (Eras)';

# Section 3

## 3.1: Bivariate Analysis of Two Qualitative Variables

SAS code for all tasks in this section.

/\*bivariate analysis\*/

proc freq data=fpcars.fpdata;

table yeargroup\*fuel\_type;

/\*visualizations\*/

\*side-by-side bar chart\*;

proc sgplot data=fpcars.fpdata;

title 'Fuel Type by Year Group';

vbar yeargroup/ group=fuel\_type groupdisplay=cluster stat=percent;

\*100% stacked bar chart\*;

proc sgplot data=fpcars.fpdata;

title 'Distribution of Fuel Type Given Year Group';

vbar yeargroup/ group=fuel\_type;

## 3.2: Bivariate Analysis of a Quantitative Variable and a Qualitative Variable

SAS code for all tasks in this section.

\*numerical summaries\*;

proc means data=fpcars.fpdata

n mean std min median max nmiss q1 q3 maxdec=3;

var kms\_driven;

class fuel\_type;

\*visualization (boxplot)\*;

proc sgplot data=fpcars.fpdata;

vbox kms\_driven / category= fuel\_type;

title "Boxplot of Kilometers Driven by Fuel Type";

yaxis label= "Number of Kilometers Driven";

\*confidence interval\*;

proc means data=fpcars.fpdata alpha=0.05 maxdec=3;

var kms\_driven;

class fuel\_type;

## 3.3: Bivariate Analysis of Two Continuous Variables

SAS code for all tasks in this section.

/\*Drawing a SRS\*/

proc surveyselect data=fpcars.fpdata out=fpcars.srs4 method=srs

sampsize=29 seed=288;

\*scatterplot\*;

proc reg data=fpcars.srs4 alpha= 0.05 plots(only)=(residuals fitplot);

model selling\_price = kms\_driven;

\*correlation\*;

proc corr data=fpcars.srs4 nomiss outp=fpcars.srs4;

var kms\_driven selling\_price;