USED CAR INDUSTRY ANALYSIS

DESCRIPTIVE ANALYSIS INTO THE 2021 USED CAR MARKET

RISHI RAJ SONI | UNIVERSITY OF SAN DIEGO | FEBRUARY 2024

TABLE OF CONTENTS

Introduction

Purpose:

- This report applies a descriptive analytics approach to answer the question of 'what exactly happened,' within the 2021 used car market. The primary purpose of this research was to combine data analysis techniques with a business mindset to discover and quantify significant relationships pertaining to the price of used cars. These findings could then be beneficial by giving those looking to potentially buy or sell a vehicle in the near future various ideas to think about or strategies to consider.

Background:

- My data comes from Kaggle, an online data science platform that hosts various competitions and collaborations for data scientists, while also providing them with real world datasets to practice solving complex issues. The author for the dataset I chose had pulled their data from craigslist car listings across the United States by using a web scraper built in Python.
- In order to conduct my own analysis, I started by cleaning the data in both Excel and R to make things easier to import into other programs while also removing unnecessary columns/rows. Once my data had been cleaned, I was able to import it into PostgreSQL and Tableau where I would be able to start building queries and visualizations to understand my data better, as well as begin finding underlying trends and key details. Data cleaning had also made working in R easier when applying regression models and other more complex predictive methods later on.

Programs/Languages:

- RStudio: Programming language used for statistical analysis, data cleaning, and visualizations.
- PostgreSQL: Free, open-source version of Microsoft SQL applied to provide detailed queries of the data.
- Tableau: Business intelligence suite used for data visualizations.
- Microsoft Excel: Spreadsheet software used in data cleaning process.

Data Insights

- Observed some of the top 5 cheapest manufactures on average.
- Looked at the top 5 most listed vehicles on the used car market.
- Created histograms to show the distribution of price and mileage in the dataset.
    - Then looked more closely at the relationship between both variables with a scatter plot.
- Analyzed how certain fuel types can have a significant effect on price and how they differ.
- Observed how different countries compare based on price and mileage.
    - Further segmented on specifically the cheapest manufactures.
- Looked at whether or not transmission type is a factor in pricing.
- Created charts to show how fast or slow model year prices have gone up by country to indicate a vehicle's long term resale value.
- Checked how color can also play a significant role in the price of a used car.

Data Cleaning

Step 1: Removing Unnecessary Columns

- Deleted columns for ID, URL, Region URL, IMG URL, Long/Lat, Country in Excel
    - ID was redundant because VIN number could be used instead.
    - The country column had all empty rows, and the data only pertained to US car listings.

Step 2: Fixing Data Format

- Used Excel to change date to YYYY-MM-DD which is the format read by SQL
- Removed Timestamps
    - Used formula =DATEVALUE(LEFT(S:S, 10)), copied values into a new column to not be stored as a formula, then replaced #NULL! values with blanks.

Step 3: Removed Extreme Values/Outliers

- Removed 51 rows containing vehicles with price and mileage values that were exceedingly high using Excel.

Step 4: Removed Rows Containing Empty VIN Numbers

- Applied R code to remove blank VIN rows, bringing down the number of total observations from 426,880 to 265,781, and reducing the file size by about 300mb.

```
#Load Libraries
library(dplyr)

#Import
df <- read.csv("C:/Users/rishi/OneDrive/Desktop/Work/Project_Portfolio/Used_Car_Project/vehicles_CLEAN.csv")
View(df)

#Filter Empty VIN
filter <- subset(df, VIN != "")
```

Step 5: Added a New Column for Country of Manufacturer

- Implemented a new column using the DPLYR package in R to assign a country value based on the manufacturer.
    - This was useful as we will see later in analyzing more trends within the data based on country.

```
#Add a Column for Country
filter <- filter %>%
  mutate(country = case_when(
    manufacturer == "bmw" | manufacturer == "mercedes-benz" | manufacturer == "porsche" | manufacturer == "audi" |
    manufacturer == "volkswagen" ~ "Germany",
    manufacturer == "honda" | manufacturer == "acura" | manufacturer == "toyota" | manufacturer == "datsun" | manufacturer
    == "nissan" | manufacturer == "infiniti" | manufacturer == "lexus" | manufacturer == "mazda" | manufacturer == "mistubishi" |
    manufacturer == "subaru" ~ "Japan",
    manufacturer == "alfa-romeo" | manufacturer == "ferrari" | manufacturer == "fiat" ~ "Italy",
    manufacturer == "aston-martin" | manufacturer == "jaguar" | manufacturer == "land rover" | manufacturer == "mini" |
    manufacturer == "rover" ~ "England",
    manufacturer == "buick" | manufacturer == "cadillac" | manufacturer == "chevrolet" | manufacturer == "chrysler" |
    manufacturer == "dodge" | manufacturer == "ford" | manufacturer == "gmc" | manufacturer == "harley-davidson" |
    manufacturer == "jeep" | manufacturer == "lincoln" | manufacturer == "mercury" | manufacturer == "pontiac" | manufacturer
    == "ram" | manufacturer == "saturn" | manufacturer == "tesla" ~ "United States",
    manufacturer == "hyundai" | manufacturer == "kia" ~ "South Korea",
    manufacturer == "volvo" ~ "Sweden",
    TRUE ~ "Empty"
  ))
```

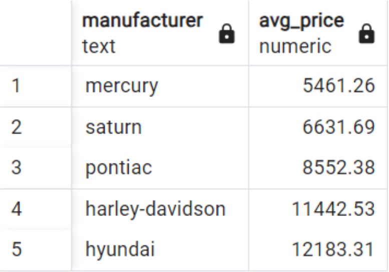
Exploratory Data Analysis

Part 1: Data Summary

- Began by querying in SQL for the least expensive car brands and the most popular cars listed. This helped provide some background on generally cheaper brands to consider, as well as what is commonly listed.

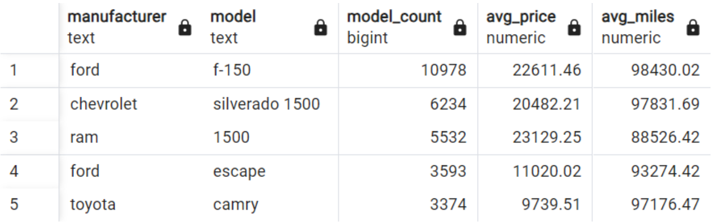
```
--Top 5 Least Expensive Car Brands
SELECT manufacturer, ROUND(AVG(price),2) AS avg_price
FROM used_cars
GROUP BY manufacturer
HAVING COUNT(manufacturer) >= 100
ORDER BY avg_price ASC
LIMIT 5;

--What are the most frequently listed cars and their prices/miles?
SELECT manufacturer, model, COUNT(model) AS model_count,
  ROUND(AVG(price),2) AS avg_price, ROUND(AVG(odometer),2) AS avg_miles
FROM used_cars
GROUP BY manufacturer, model
ORDER BY model_count DESC
LIMIT 5;
```

Table 1



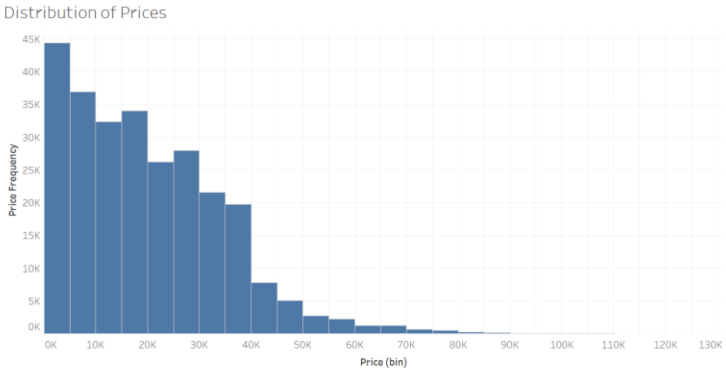
	manufacturer text	avg_price numeric
1	mercury	5461.26
2	saturn	6631.69
3	pontiac	8552.38
4	harley-davidson	11442.53
5	hyundai	12183.31

Table 2



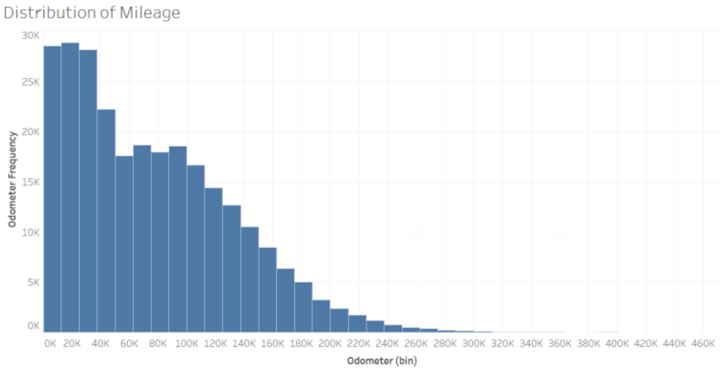
	manufacturer text	model text	model_count bigint	avg_price numeric	avg_miles numeric
1	ford	f-150	10978	22611.46	98430.02
2	chevrolet	silverado 1500	6234	20482.21	97831.69
3	ram	1500	5532	23129.25	88526.42
4	ford	escape	3593	11020.02	93274.42
5	toyota	camry	3374	9739.51	97176.47

- Next, I created two histograms in Tableau to show the distribution of price and odometer data. This helped later when plotting the relationship between the two variables because I could filter for where a rough majority of the data lies to exclude outliers.



The trend of count of Price for Price (bin). The data is filtered on Price (bin), which has multiple members selected.

Figure 1



The trend of count of Odometer for Odometer (bin). The data is filtered on Odometer (bin), which has multiple members selected.

Figure 2

- From Figure 1, the histogram shows most of our price data falls between $0-5,000 and is positively skewed. Indicating most prices are on the lower end, but there are some higher values stretching the data out to the right. Similarly for Figure 2 and odometer, there is another positive skew of the data, and much of our odometer values lie around 0-40,000 miles.

Part 2: Data Analysis

- I began my deep dive into analyzing trends in the data by creating a scatterplot using the ggplot2 package within R to examine the relationship between price and odometer values. From this plot, we can see that price is no longer impacted by mileage past roughly 150,000. This indicated that perhaps purchasing vehicles past 150,000 miles will not yield the best value for money. When selling, it is best to keep the vehicle under 100,000 miles.

```
#A Look at Price and Mileage
df_filtered <- df %>%
  filter(odometer <= 300000, price <= 100000)

ggplot(df_filtered, aes(x = odometer, y = price)) + geom_point() +
  labs(title = "Relationship Between Mileage and Price", x = "Miles", y = "Price") +
  scale_x_continuous(labels = label_number()) +
  scale_y_continuous(labels = label_number()) +
  geom_smooth(method = "lm", formula = y ~ poly(x, 2), color = 'blue')
```

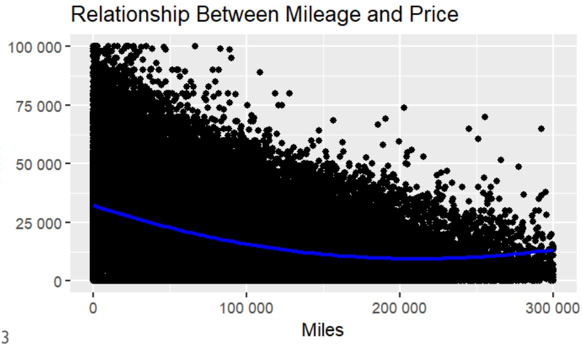
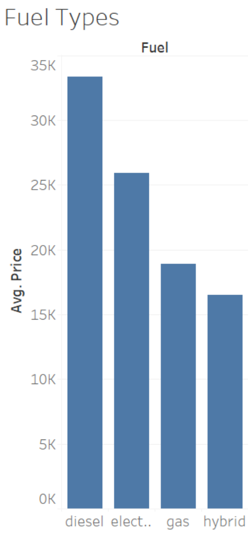


Figure 3

- Next, I built a basic linear model, conducting a regression analysis, to assess the impact of certain fuel types on price, and check for statistical significance. This was done by using R's lm() function.

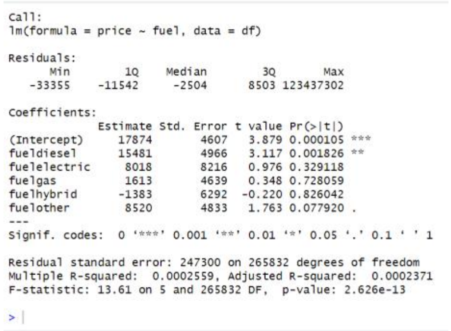
```
#Build a Linear Model for Types
M1 <- lm(price ~ fuel, df)
summary(M1)
```

- This model showed us high statistical significance, or low p-values, for diesels having a very high impact on price. To further investigate this, I then observed the average prices broken down by fuel types in SQL and then put it into a visual with Tableau.



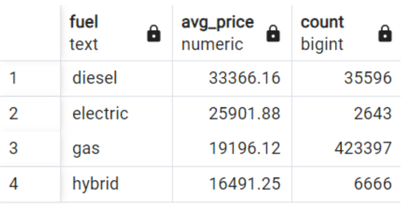
Average of Price for each Fuel. The view is filtered on Fuel, which excludes Null and other.

Figure 4



```
Call:
lm(formula = price ~ fuel, data = df)

Residuals:
     Min       1Q   Median       3Q       Max
  -33355   -11542    -2504     8503 123437302

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     17874       4607   3.879 0.000105 ***
fueldiesel      15481       4966   3.117 0.001826 **
fuelelectric     8018       8216   0.976 0.329118
fuelgas          1613       4639   0.348 0.728059
fuelhybrid      -1383       6292  -0.220 0.826042
fuelother        8520       4833   1.763 0.077920 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 247300 on 265832 degrees of freedom
Multiple R-squared:  0.0002559, Adjusted R-squared:  0.0002371
F-statistic: 13.61 on 5 and 265832 DF,  p-value: 2.626e-13

> |
```

Figure 5



	fuel text	avg_price numeric	count bigint
1	diesel	33366.16	35596
2	electric	25901.88	2643
3	gas	19196.12	423397
4	hybrid	16491.25	6666

Table 3

- From these tables and figures, we overall see that diesels are the most expensive vehicle by a large margin. While the data does indicate hybrids as the cheapest option, it is important to consider the low sample size of hybrids in this particular dataset, and it is not labeled as overall being a statistically significant variable.

- I then wanted to observe how different countries' vehicles compare in price as well as mileage with one another. For this analysis I created a bar chart with Tableau to visualize average price by country based on odometer.

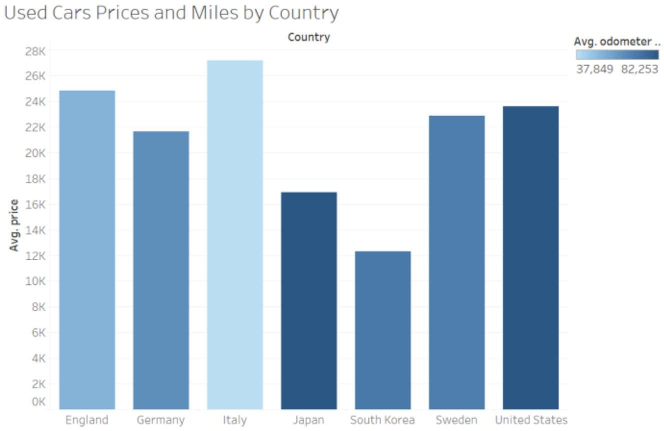

Used Cars Prices and Miles by Country

Figure 5

Average of price (Used Cars CLEAN.csv) for each Country. Color shows average of odometer (Used Cars CLEAN.csv). The view is filtered on Country, which excludes Empty.

- From this chart, we can quickly gather that South Korean makes tend to be significantly cheaper than other countries while also maintaining a respectable odometer value as well. Meanwhile, US car manufacturers seem to price their cars on the higher end while still having a high mileage count. When compared to Japanese cars, that offer similar mileage on average to the US but with much lower prices, they also emerge as a top contender for used cars. Even German cars had on average been cheaper with less miles than the US.

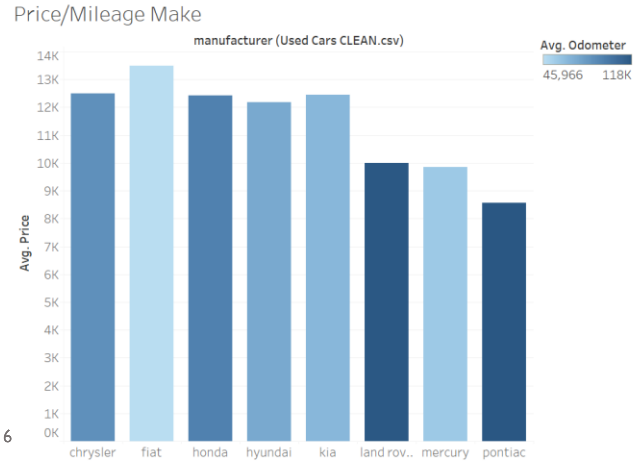

Price/Mileage Make

Figure 6

Average of Price for each manufacturer (Used Cars CLEAN.csv). Color shows average of Odometer. The data is filtered on Manufacturer, which keeps 34 of 42 members. The view is filtered on manufacturer (Used Cars CLEAN.csv), which keeps 10 of 42 members.

- I then decided to segment this further, looking specifically at the 8 cheapest manufactures based on the same parameters. As expected, KIA and Hyundai from South Korea were front running options. However, we also saw Fiat and Mercury appear as strong options as well. While Fiat's were slightly more expensive, they had the lowest odometer value of 45,966 on average. Mercury too are low in miles at an affordable price, however it is worth keeping in mind the brand has been discontinued, and this may factor into its pricing.

- The next metric I was interested in looking at was whether or not transmission type played a role in impacting price. I began by querying for average automatic vs manual prices in SQL and found overall comparable results with automatics being marginally higher. I then created a line chart in Tableau to break it down by model year to check if it's better to buy automatic or manual on a newer vs older car.

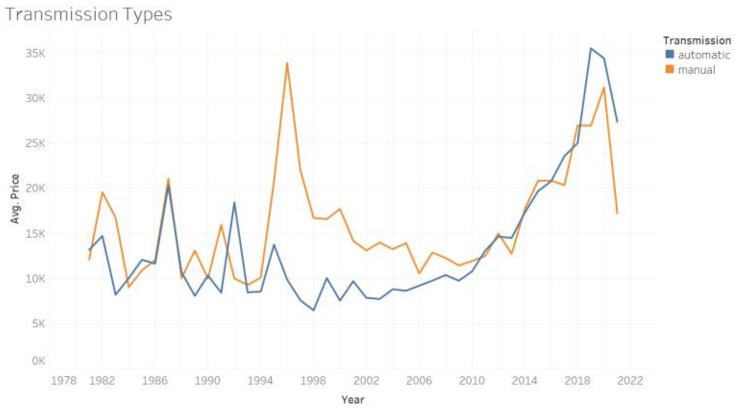

Transmission Types

Figure 7

The trend of average of Price for Year. Color shows details about Transmission. The view is filtered on Transmission and Year. The Transmission filter excludes Null and other. The Year filter ranges from 1981 to 2021.

- The chart overall showed that automatic and manual cars are similar in terms of price. There was however a spike from 1994-2010 where manuals were far more expensive. I believe this can be attributed to automatics being still less common and refined of a technology until the late 2000s and manuals could be more desirable on older cars or desirables of that era.

Part 2: Data Analysis (continued)

- I then looked at how model years from different countries have changed in price over the last 20 years. Generally, as you would expect, newer models steadily increase in price. However, we notice that some country's older vehicles depreciate faster. This was most prevalent in German cars and with the US as well. Whether you're buying or selling for the long term, it is insightful to know the resale value of your vehicle.

```
#Observe Average Prices for Countries over the last 20 years
df_filtered4 <- df[df$country %in% c("United States", "Japan", "Germany", "South Korea") & df$year >= 2001 & df$year <= 2021, ]

ggplot(df_filtered4, aes(x = year, y = price)) +
  geom_point(stat = "summary", fun = "mean", color = "red") +
  scale_x_continuous(breaks = seq(2001, 2021, by = 1)) +
  labs(title = "Average Price Over the Last 20 Years",
       x = "Year",
       y = "Average Price",
       color = "Country") +
  facet_wrap(~country) +
  theme(axis.text.x = element_text(size = 4))
```

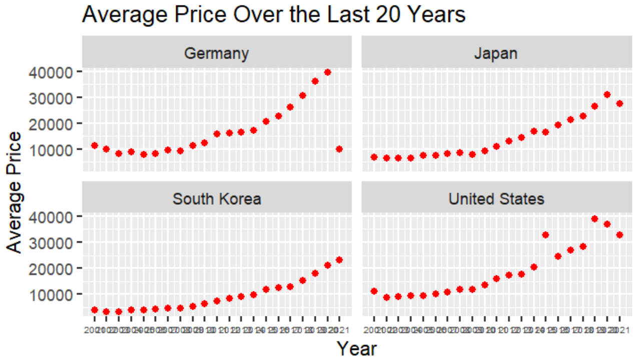
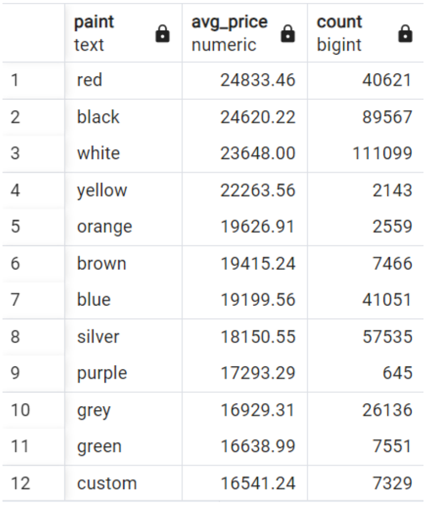


Figure 8

- For my final observation, I looked at how paint color can also have an impact on price by creating another query in SQL. Interestingly, I found there was a significant amount of grey and custom paint jobs that had much lower prices on average than other colors.

```
--Avg Prices based on Color
SELECT paint, ROUND(AVG(price),2) AS avg_price, COUNT(paint)
FROM used_cars
WHERE paint != ‘’
GROUP BY paint
ORDER BY avg_price DESC;
```



	paint text	avg_price numeric	count bigint
1	red	24833.46	40621
2	black	24620.22	89567
3	white	23648.00	111099
4	yellow	22263.56	2143
5	orange	19626.91	2559
6	brown	19415.24	7466
7	blue	19199.56	41051
8	silver	18150.55	57535
9	purple	17293.29	645
10	grey	16929.31	26136
11	green	16638.99	7551
12	custom	16541.24	7329

Table 4

- With a rather significant sample size for colors, it was fascinating to see how big of an impact it could play onto price. Red cars were on average a whole $8,000 more expensive than grey or custom paint jobs. It may be worth noting to buyers or sellers when considering repainting a vehicle, it can also have a considerable effect on its value.

Conclusion

Key Findings:

- The effect of mileage on price is generally not impactful past 150,000 miles. Purchasing past this milage typically won't yield the best value for money.
- Diesels tended to be the most expensive cars on average, exceeding even electrics. Hybrids were cheapest, although not statistically as significant.
- Japanese and South Korean makes tended to have the best value when it came to the cheapest prices and lowest miles. When looking into specific manufactures, KIA, Hyundai, Mercury, and Fiat emerged as the top options.
- Transmission type generally is not a factor, unless potentially looking at select older vehicles.
- When observing the average price by model year for different countries, some countries depreciate notably faster which is worth keeping note of for long term purchases.
- Grey cars, as well as custom paint jobs had a significantly lower price on average than other colors.

Further Research:

- When continuing exploring this project, some points I would have liked to observe would have been having at least three more columns for the original vehicle's MSRP, accident history, as well as the number of previous owners. MSRP could have been a useful and perhaps more accurate metric when analyzing how vehicles depreciate over time than compared to checking model years value. Accident history could also help give better insight on how value is impacted by the number of accidents and how severe they were. And the number of previous owners would have been useful as well to analyze price, generally more owners could mean the vehicle had been neglected by having many owners in a short period of time.
- The last exploratory research experiment I would have conducted would have been to create a predictive model by applying the statistics and the information gained from different models in R. A function, or algorithm, to predict the value of a used car based on parameters such as miles, manufacturer, or year, input by a user.

Recommendations:

- When looking at purchasing a used vehicle, the top countries to consider based on average price and miles are Japan and South Korea. Both countries tend to offer prices on the lower end while still not being egregiously high on mileage either. Delving into specific manufactures, these would be KIA and Hyundai, but if you are willing to spend slightly more, Fiat is a great option as well. When looking to grasp the best value for money, searching for vehicles above 150,000 miles will be redundant as prices no longer see any change, and so it is best to filter your results below this. Diesel cars are the most expensive buying option, and so it may be worth staying away from these in your searches. Hybrids were on average the cheapest, however this may not always be the case as the sample size for this data set was relatively small. When it came to your car's resale value, the best long-term options were found in Japanese and South Korean makes again which held their value the most consistent across model years over time. German cars saw the largest amount of price increase with newer models, followed by the US. Transmission type will generally not impact your search as they were all around found to be similar in price, with potential discrepancies in the late 1990s and early 2000s with manuals costing more. Lastly, the best bargain on paint jobs are grey and custom paints, which showed considerable price decreases than compared to other colors.