`Car Price Prediction Based on Consumer Traits`

STAT 301 Project Group 22 Project Proposal

By Janice Chan, Elias Khan, Davis Li, Daniel Yuan

Introduction

One of our team members, Davis, is studying business and works at a Honda retailer selling cars. We found this car dataset from Kaggle that stood out to us because it is directly correlated to what we're interested in. As a salesman, it is important to gauge how much a client is willing to pay for a car to make the best suggestions. This dataset helps us answer that question by providing input variables (demographics/attributes of a buyer) to predict the response variable (the price they paid for the car).

Question:

The key question we inquire is: "what attributes of a buyer can best predict how much a buyer is willing to pay for a car?". As our project is based on prediction, we want to predict: "based on attributes of a buyer/consumer, how much would they be willing to pay for a car?".

Dataset:

"Car Sales Price Prediction" from Kaggle: https://www.kaggle.com/datasets/yashpaloswal/ann-car-sales-price-prediction (Yashpal, n.d.), consisting of 500 observations.

The input variables are customer name, customer email, country, gender, age, annual salary, credit card debt, and net worth.

The response variable is the amount paid for a car by the buyer.

A drawback to our dataset is that it does not indicate the currency type for the variables regarding currency and whether they differ, which may conflict as there are various countries. However, we chose to generalize that they are all the same currency as listed together in one variable for convenience and decided to filter out countries in response to possible confusion.

Relevant Research

From previous studies, we know that the attributes of buyers affect how much they spend on a car. A scientific study by Chandra et al. (2013) describes that older buyers spend more on a car; a trend particularly illustrated in women, indicating that the gender and age of the buyer affect how much they pay for the car. Another study found that income influenced the choice of car price; not gender or age (Rimple et al., 2015).

However, the study by Chandra et al. (2013) does not consider consumers' income/wealth, and the study by Rimple et al. (2015) was limited to a small sample size of 164 respondents in India. Thus, our research will expand on previous studies to gain a better understanding of how customer's traits may affect how much a buyer would spend on a car with a larger and more diverse dataset; providing a better model for prediction and knowledge of the relationship between buyers and their car purchases.

Methods

Preliminary analysis

Before we begin building our prediction model, we perform some preliminary analysis.

The first step is to load in the libraries we will need to analyse the dataset and setting the seed for reproducability:

```
set.seed(1234)

library(tidyverse)
library(repr)
library(digest)
library(infer)
library(broom)
library(leaps)
library(mltools)
library(glmnet)
library(httr)

options(juptyer.plot_mimetypes = "image/png")
```

```
── Attaching packages ───────────────────────────────────────
tidyverse 1.3.2 ──
✔ ggplot2 3.3.6     ✔ purrr   0.3.5
✔ tibble  3.1.8     ✔ dplyr   1.0.10
✔ tidyr   1.2.1     ✔ stringr 1.4.1
✔ readr   2.1.3     ✔ forcats 0.5.2
── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()

Attaching package: ‘mltools’


The following object is masked from ‘package:tidyr’:

    replace_na


Loading required package: Matrix


Attaching package: ‘Matrix’


The following objects are masked from ‘package:tidyr’:

    expand, pack, unpack
```

```
Loaded glmnet 4.1-2
```

Now we can load in the dataset:

```
url <- "https://drive.google.com/uc?
export=download&id=1smVyESJZSdTi6EeQBs7g09cZSmuz-CmE"
raw_car_data <- read_csv(url)
head(raw_car_data)
```

```
Rows: 500 Columns: 9
── Column specification
────────────────────────────────────────────────────────
Delimiter: ","
chr (3): customer name, customer e-mail, country
dbl (6): gender, age, annual Salary, credit card debt, net worth, car
purcha...

ⓘ Use `spec()` to retrieve the full column specification for this
data.
ⓘ Specify the column types or set `show_col_types = FALSE` to quiet
this message.

  customer name   customer e-mail

1 Martina Avila   cubilia.Curae.Phasellus@quisaccumsanconvallis.edu

2 Harlan Barnes   eu.dolor@diam.co.uk

3 Naomi Rodriquez
vulputate.mauris.sagittis@ametconsectetueradipiscing.co.uk
4 Jade Cunningham malesuada@dignissim.com

5 Cedric Leach    felis.ullamcorper.viverra@egetmollislectus.net

6 Carla Hester    mi@Aliquamerat.edu

  country      gender age      annual Salary credit card debt net
worth
1 Bulgaria     0      41.85172 62812.09      11609.381        238961.3

2 Belize       0      40.87062 66646.89       9572.957        530973.9

3 Algeria      1      43.15290 53798.55      11160.355        638467.2

4 Cook Islands 1      58.27137 79370.04      14426.165        548599.1

5 Brazil       1      57.31375 59729.15       5358.712        560304.1

6 Liberia      1      56.82489 68499.85      14179.472        428485.4
```

```
  car purchase amount
1 35321.46
2 45115.53
3 42925.71
4 67422.36
5 55915.46
6 56612.00
```

Before we begin analyzing the data, it's important to decide which columns are important for our research and to clean the data if necessary.

Our first course of action will be to change the names of columns so that there are no spaces or symbols like "-" that will hinder our ability to analyze the data.

```
colnames(raw_car_data) <- gsub(" ", "_", colnames(raw_car_data))
colnames(raw_car_data) <- gsub("-", "", colnames(raw_car_data))
head(raw_car_data)
```

```
  customer_name   customer_email

1 Martina Avila   cubilia.Curae.Phasellus@quisaccumsanconvallis.edu

2 Harlan Barnes   eu.dolor@diam.co.uk

3 Naomi Rodriquez
vulputate.mauris.sagittis@ametconsectetueradipiscing.co.uk
4 Jade Cunningham malesuada@dignissim.com

5 Cedric Leach    felis.ullamcorper.viverra@egetmollislectus.net

6 Carla Hester    mi@Aliquamerat.edu

  country      gender age      annual_Salary credit_card_debt
net_worth
1 Bulgaria     0      41.85172 62812.09      11609.381        238961.3

2 Belize       0      40.87062 66646.89       9572.957        530973.9

3 Algeria      1      43.15290 53798.55      11160.355        638467.2

4 Cook Islands 1      58.27137 79370.04      14426.165        548599.1

5 Brazil       1      57.31375 59729.15       5358.712        560304.1

6 Liberia      1      56.82489 68499.85      14179.472        428485.4

  car_purchase_amount
1 35321.46
2 45115.53
3 42925.71
```

```
4 67422.36
5 55915.46
6 56612.00
```

Next we must change the gender column so that it is categorical, rather than numerical.

```
raw_car_data$gender[raw_car_data$gender == 0] <- "Male"
raw_car_data$gender[raw_car_data$gender == 1] <- "Female"
raw_car_data$gender = as.factor(raw_car_data$gender)

head(raw_car_data)

  customer_name   customer_email

1 Martina Avila   cubilia.Curae.Phasellus@quisaccumsanconvallis.edu

2 Harlan Barnes   eu.dolor@diam.co.uk

3 Naomi Rodriquez
vulputate.mauris.sagittis@ametconsectetueradipiscing.co.uk
4 Jade Cunningham malesuada@dignissim.com

5 Cedric Leach    felis.ullamcorper.viverra@egetmollislectus.net

6 Carla Hester    mi@Aliquamerat.edu

  country      gender age      annual_Salary credit_card_debt
net_worth
1 Bulgaria     Male   41.85172 62812.09      11609.381        238961.3

2 Belize       Male   40.87062 66646.89       9572.957        530973.9

3 Algeria      Female 43.15290 53798.55      11160.355        638467.2

4 Cook Islands Female 58.27137 79370.04      14426.165        548599.1

5 Brazil       Female 57.31375 59729.15       5358.712        560304.1

6 Liberia      Female 56.82489 68499.85      14179.472        428485.4

  car_purchase_amount
1 35321.46
2 45115.53
3 42925.71
4 67422.36
5 55915.46
6 56612.00
```

Finally, we must remove the columns "customer name", "customer e-mail", and "country", as they will be not included in our research.

These columns vary far too much and it would be nearly impossible to make any meaningful remarks from them. Name and e-mail are particular customers, to which we do not need to consider as we are not identifying for anyone. For this particular case, we did not include country into consideration given that there were many different countries but the dataset was quite small (resulting in each country not having that many instances). For example, the most counts of a country's is 6 for (Israel, Mauritania, and Bolivia).

```
car_data <- raw_car_data %>% select(-customer_name, -customer_email, -
country)
head(car_data)
```

```
  gender age      annual_Salary credit_card_debt net_worth
car_purchase_amount
1 Male   41.85172 62812.09      11609.381        238961.3  35321.46

2 Male   40.87062 66646.89       9572.957        530973.9  45115.53

3 Female 43.15290 53798.55      11160.355        638467.2  42925.71

4 Female 58.27137 79370.04      14426.165        548599.1  67422.36

5 Female 57.31375 59729.15       5358.712        560304.1  55915.46

6 Female 56.82489 68499.85      14179.472        428485.4  56612.00
```

Now we have clean data with only categorical and numerical values, we can begin analyzing the data.

We will first create a distribution of car purchase amounts to see how the amount people spend on cars varies overall; looking at the summary statistics.

```
mean_purchases = mean(car_data$car_purchase_amount) %>%
    round(digits = 2)
median_purchases = median(car_data$car_purchase_amount) %>%
    round(digits = 2)
sd_purchases = sd(car_data$car_purchase_amount) %>%
    round(digits = 2)
```

Table 1. Data Summary of Estimates (rounded)

Mean car purchase amount	Median car purchase amount	Standard Deviation of car purchase amounts
$44209.80	$43997.78	$10773.18

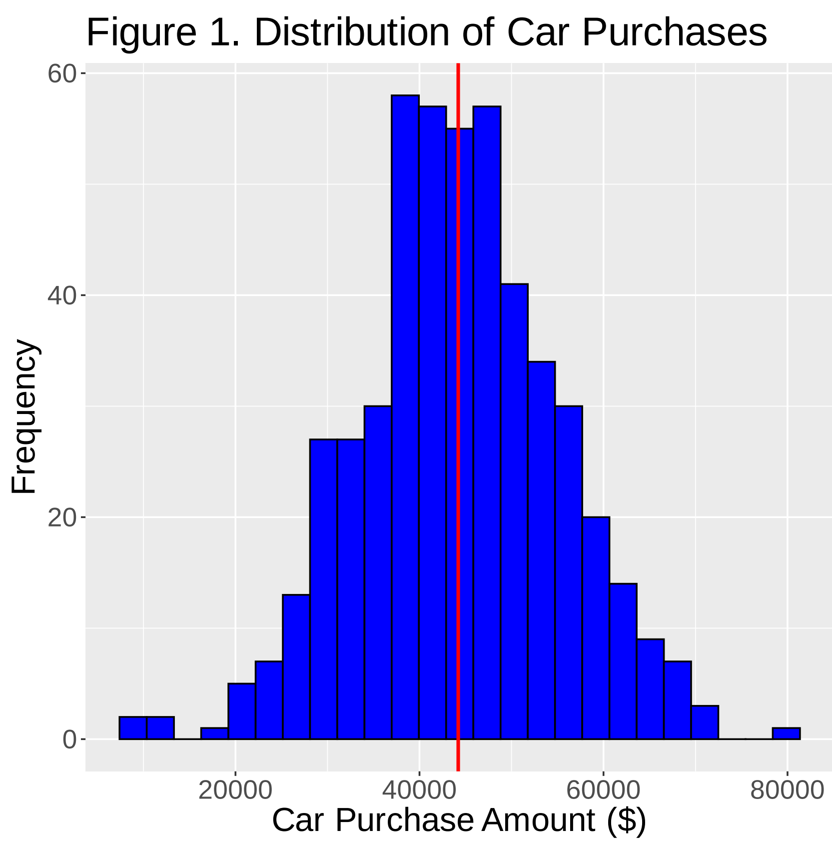
```
quantile_purchases = quantile(car_data$car_purchase_amount) %>%
    round(digits = 2)
```

Table 2. Quantiles of Car Purchases (rounded)

0%	25%	50%	75%	100%
9000	37629.90	43997.78	51254.71	80000

We create various plots exploring the relationship of our response variable (car purchase amount) with the other variables in our dataset.

```
mean_purchase = mean(car_data$car_purchase_amount)

car_data %>% ggplot(aes(x=car_purchase_amount)) +
    geom_histogram(bins = 25, color = "black", fill = "blue") +
    geom_vline(xintercept= mean_purchase, color = "red", size = 1) +
    xlab("Car Purchase Amount ($)") +
    ylab("Frequency") +
    ggtitle("Figure 1. Distribution of Car Purchases") +
    theme(text = element_text(size = 20))
```



These statistics combined with the visualisation above give us some valuable information about our data:

1. The distribution of car purchases is relatively symmetrical. We can see this visually, and we can confirm this by observing that the mean (44209.80 dollars) and the median (43997.78 dollars) are very close in value. This tells us that there are

roughly the same amount of people that spend more than the mean than those who spend less.

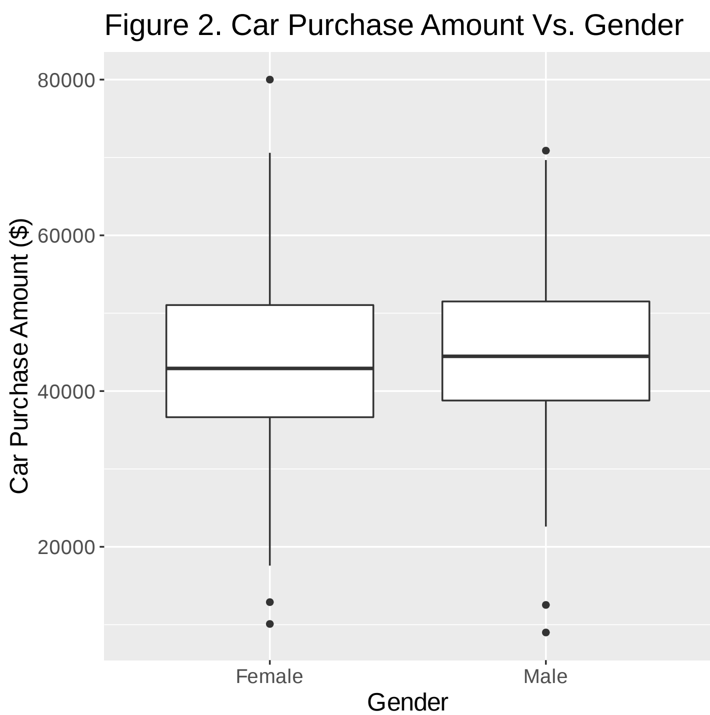
2. 50% of the purchases lie between 37629.90 and 51254.71 dollars. This indicates to us that the amount that customers spend doesn't vary that much since the upper and lower quartile are relatively close. This is also shown in how high the peak of the distribution is compared to the outer sections of the distribution.

Now that we've become a bit more familiar with our response variable, we can begin seeing how other variables in the dataset compare and correlate with our response variable.

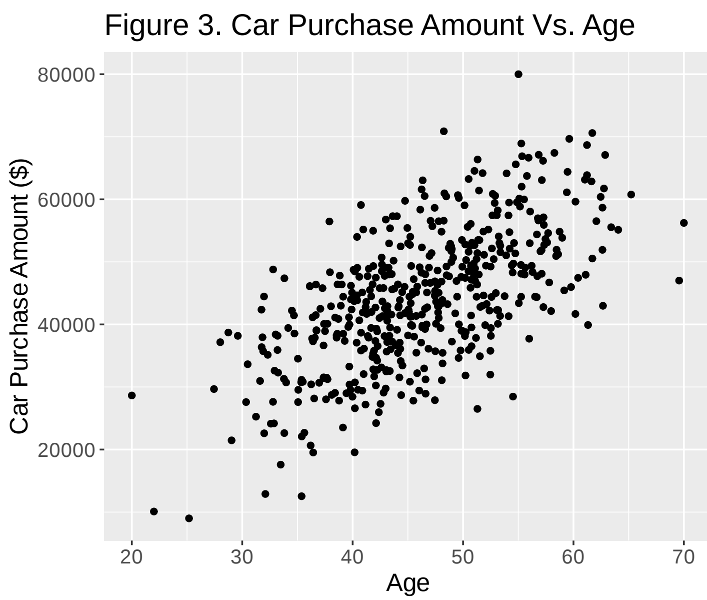
```
head(car_data)
```

	gender	age	annual_Salary	credit_card_debt	net_worth	car_purchase_amount
1	Male	41.85172	62812.09	11609.381	238961.3	35321.46
2	Male	40.87062	66646.89	9572.957	530973.9	45115.53
3	Female	43.15290	53798.55	11160.355	638467.2	42925.71
4	Female	58.27137	79370.04	14426.165	548599.1	67422.36
5	Female	57.31375	59729.15	5358.712	560304.1	55915.46
6	Female	56.82489	68499.85	14179.472	428485.4	56612.00

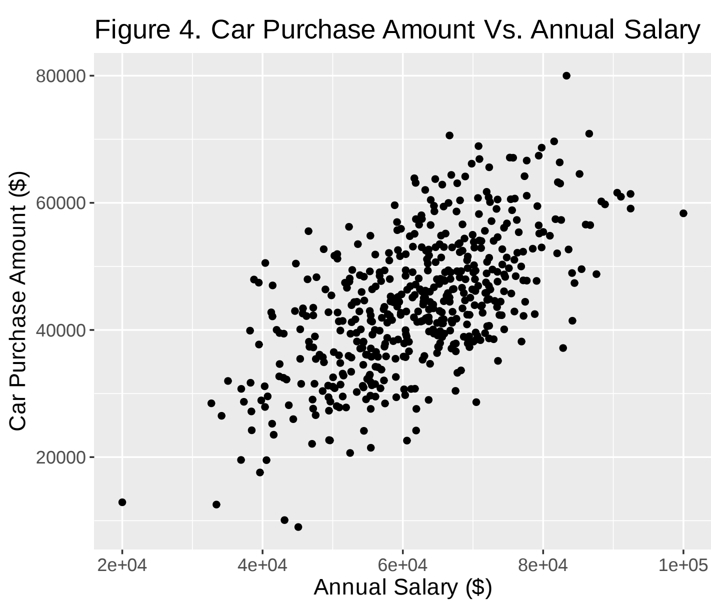
```
options(repr.plot.width = 6, repr.plot.height = 6)

gender_plot <- car_data %>% ggplot(aes(x = gender, y = 
car_purchase_amount)) +
                    geom_boxplot() +
    xlab("Gender")+
    ylab("Car Purchase Amount ($)")+
    ggtitle("Figure 2. Car Purchase Amount Vs. Gender") +
    theme(text = element_text(size = 15))


gender_plot
```



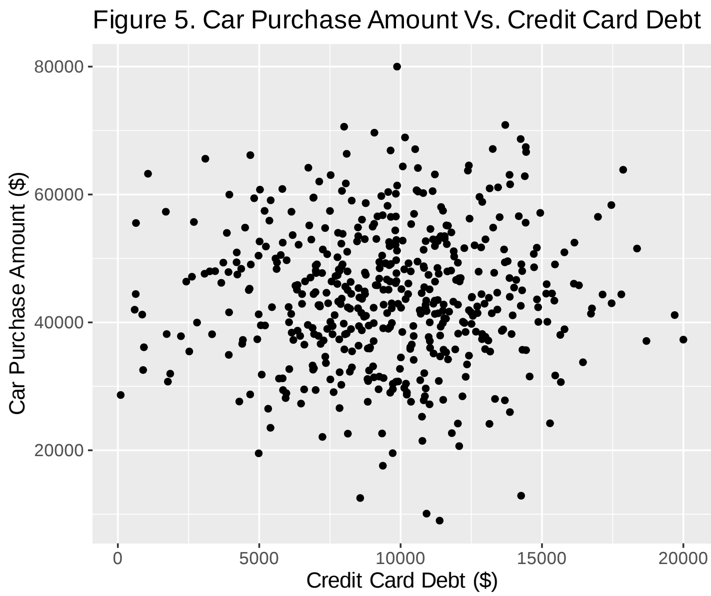

```
options(repr.plot.width = 6, repr.plot.height = 5)

age_plot <- car_data %>% ggplot(aes(x = age, y = car_purchase_amount))
+
                    geom_point() +
    xlab("Age")+
    ylab("Car Purchase Amount ($)")+
    ggtitle("Figure 3. Car Purchase Amount Vs. Age") +
    theme(text = element_text(size = 15))

age_plot
```



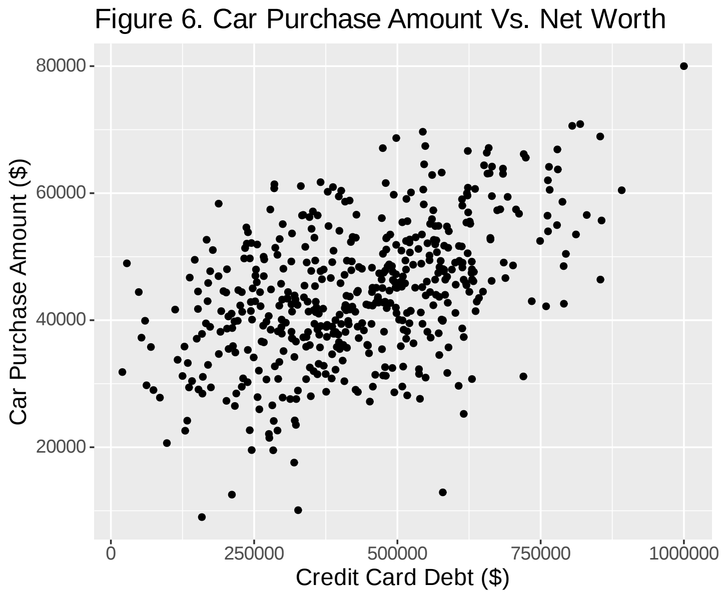

```
options(repr.plot.width = 6, repr.plot.height = 5)

annual_Salary_plot <- car_data %>% ggplot(aes(x = annual_Salary, y =
car_purchase_amount)) +
                    geom_point() +
    xlab("Annual Salary ($)")+
    ylab("Car Purchase Amount ($)")+
    ggtitle("Figure 4. Car Purchase Amount Vs. Annual Salary") +
    theme(text = element_text(size = 13.5))

annual_Salary_plot
```



```
options(repr.plot.width = 6, repr.plot.height = 5)

credit_card_debt_plot <- car_data %>% ggplot(aes(x = credit_card_debt,
y = car_purchase_amount)) +
                    geom_point() +
    xlab("Credit Card Debt ($)")+
    ylab("Car Purchase Amount ($)")+
    ggtitle("Figure 5. Car Purchase Amount Vs. Credit Card Debt") +
    theme(text = element_text(size = 13))

credit_card_debt_plot
```




```
options(repr.plot.width = 6, repr.plot.height = 5)

net_worth_plot <- car_data %>% ggplot(aes(x = net_worth, y =
car_purchase_amount)) +
                    geom_point() +
    xlab("Credit Card Debt ($)")+
    ylab("Car Purchase Amount ($)")+
    ggtitle("Figure 6. Car Purchase Amount Vs. Net Worth") +
    theme(text = element_text(size = 14))

net_worth_plot
```



After visualizing how different variables stack against car purchase amount, it is safe to say that we will definitely be able to make a solid predictive model for our response variable. More than one of these plot figures shows a strong positive correlation to the response variable, which gives us the confidence to end our exploratory analysis of the data and begin preparing to fully analyze it and build models.

Analysis and Model Building

We start by splitting our dataset into a training and testing set; creating the training set by sampling without replacement with 60% of observations in our dataset and using anti_join() and 'ID' to create our testing set. This way, we can generate a model from the training set, and use it to predict values in the test set to test our model's prediction performance.

```
car_data$ID <- 1:nrow(car_data)
training_car <- sample_n(car_data, size = nrow(car_data) * 0.60,
replace = FALSE)

testing_car <- anti_join(car_data, training_car, by = "ID")

head(training_car, 3)
head(testing_car, 3)

  gender age      annual_Salary credit_card_debt net_worth
car_purchase_amount
1 Female 40.87537 59060.09       5841.612        136346.31 29417.65

2 Male   61.31742 51086.88      12254.539         59630.08 39911.61

3 Male   45.75423 63172.96       6332.202        456524.79 45112.95

  ID
1 284
```

```
2 336
3 406

  gender age      annual_Salary credit_card_debt net_worth
car_purchase_amount
1 Male   41.85172 62812.09      11609.381        238961.3  35321.46

2 Female 43.15290 53798.55      11160.355        638467.2  42925.71

3 Female 57.31375 59729.15       5358.712        560304.1  55915.46

  ID
1 1
2 3
3 5
```

Since ID doesn't serve any further purpose (just helped create our test set), we will remove it.

```
training_car <- training_car %>% select(-ID)
testing_car <- testing_car %>% select(-ID)
```

Then we estimate an additive MLR with all input variables in our cleaned dataset.

```
car_full_OLS <- lm(car_purchase_amount ~ ., data = training_car)
print("Table 3. Full OLS Linear Model")
tidy(car_full_OLS)

[1] "Table 3. Full OLS Linear Model"

  term             estimate      std.error    statistic     p.value
1 (Intercept)      -4.214698e+04 7.146126e-01 -5.897878e+04 0.0000000
2 genderMale        9.911072e-02 1.702344e-01  5.822016e-01 0.5608771
3 age               8.415575e+02 1.017777e-02  8.268581e+04 0.0000000
4 annual_Salary     5.623342e-01 7.383222e-06  7.616379e+04 0.0000000
5 credit_card_debt -1.051798e-05 2.445774e-05 -4.300470e-01 0.6674765
6 net_worth         2.898336e-02 4.928031e-07  5.881327e+04 0.0000000
```

We then obtain the out-of-sample predictions for the testing set.

```
car_test_pred_full_OLS <- predict(car_full_OLS, newdata = testing_car)
head(car_test_pred_full_OLS)

       1        2        3        4        5        6
35320.91 42926.24 55913.02 28924.15 47434.85 48011.75
```

Now we calculate the root mean squared error (RMSE) for the above predictions with respect to the testing set's observed car purchase amount ($). We can compare this later to our selected model in order to ensure we get a better fit.

```
car_R_MSE_models <- tibble(
  Model = "OLS Full Regression",
  R_MSE = rmse(
    preds = car_test_pred_full_OLS,
    actuals = testing_car$car_purchase_amount))
print("Table 4. Root mean squared error")
car_R_MSE_models
```

```
[1] "Table 4. Root mean squared error"
```

```
  Model               R_MSE
1 OLS Full Regression 1.520315
```

Then we run the forward stepwise selection algorithm for the models for 1 to 5 input variables to determine which variables to choose from this automated model selection that considers all model possibilities for all sizes.

```
set.seed(1234)
car_forward_sel <- regsubsets(
  x = car_purchase_amount ~ ., nvmax = 5,
  data = training_car,
  method = "forward",
)

car_forward_summary <- summary(car_forward_sel)
car_forward_summary
```

```
Subset selection object
Call: regsubsets.formula(x = car_purchase_amount ~ ., nvmax = 5, data
= training_car,
    method = "forward", )
5 Variables  (and intercept)
                 Forced in Forced out
genderMale           FALSE      FALSE
age                  FALSE      FALSE
annual_Salary        FALSE      FALSE
credit_card_debt     FALSE      FALSE
net_worth            FALSE      FALSE
1 subsets of each size up to 5
Selection Algorithm: forward
         genderMale age annual_Salary credit_card_debt net_worth
1  ( 1 ) " "        "*" " "           " "              " "
2  ( 1 ) " "        "*" "*"           " "              " "
3  ( 1 ) " "        "*" "*"           " "              "*"
4  ( 1 ) "*"        "*" "*"           " "              "*"
5  ( 1 ) "*"        "*" "*"           "*"              "*"
```

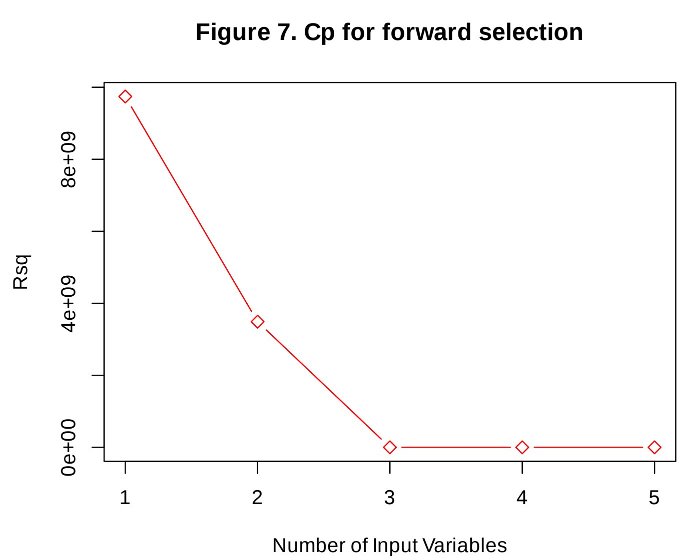
We evaluate the metrics in order to select the model above by goodness of fit.

```
car_forward_summary_df <- tibble(
    n_input_variables = 1:5,
    RSQ = car_forward_summary$rsq,
    RSS = car_forward_summary$rss,
    ADJ.R2 = car_forward_summary$adjr2,
    Cp = car_forward_summary$cp,
    BIC = car_forward_summary$bic,
)
car_forward_summary_df
```

	n_input_variables	RSQ	RSS	ADJ.R2	Cp	BIC
1	1	0.4393132	2.088299e+10	0.4374317	9.742405e+09	-162.1703
2	2	0.7992297	7.477767e+09	0.7978777	3.488553e+09	-464.5668
3	3	1.0000000	6.313310e+02	1.0000000	2.530700e+00	-5345.0739
4	4	1.0000000	6.305898e+02	1.0000000	4.184940e+00	-5339.7225
5	5	1.0000000	6.301934e+02	1.0000000	6.000000e+00	-5334.2074

We plot the Cp values for the models in our forward selection algorithm, illustrating how it changes when the model changes by adding another variable.

```
plot(summary(car_forward_sel)$cp,
  main = "Figure 7. Cp for forward selection",
  xlab = "Number of Input Variables", ylab = "Rsq", type = "b", pch =
5,
  col = "red")
```



The R2 and adjusted R2 values increase as there are more variables included in the model. We notice that the Cp statistic is decreasing until 3 variables. Using this, we select a predictive model with 3 variables, which are age, annual salary, and net worth by our models fit from the forward selection algorithm.

With our selected model, we fit it with the training set to then predict the data in the testing set. Showing its predictive performance.

```
# Estimation
car_red_OLS <- lm(car_purchase_amount ~ age + annual_Salary +
net_worth,
  data = training_car)

# Prediction
car_test_pred_red_OLS <- predict(car_red_OLS, newdata = testing_car[,
1:5])
head(car_test_pred_red_OLS)
```

```
       1        2        3        4        5        6
35320.87 42926.30 55913.03 28924.16 47434.92 48011.63
```

We compute the new RMSE of this new predictive model using the testing set and add it with that of the additive model (with all input variables) to compare and indicate if our selected model is a better fit for a better predictive performance; better prediction.

```
car_R_MSE_models <- rbind(
  car_R_MSE_models,
  tibble(
    Model = "OLS Reduced Regression",
    R_MSE = rmse(
      preds = car_test_pred_red_OLS,
      actuals = testing_car$car_purchase_amount)))
print("Table 5. Root mean squared error")
car_R_MSE_models
```

```
[1] "Table 5. Root mean squared error"
```

```
  Model                  R_MSE
1 OLS Full Regression    1.520315
2 OLS Reduced Regression 1.519226
```

Predictions are random variables, so they have uncertainty. Thus, we look at confidence intervals that account for the sample-to-sample variation of the predictions. Two such confidence intervals exist: confidence intervals for prediction and prediction intervals. They are similar but ultimately different, the former measures the expected value of the response given our prediction variables while the other measures the actual value. We look at both.

```
car_cip <- testing_car   %>%
   select(car_purchase_amount, age, annual_Salary, net_worth) %>%
   cbind(predict(car_red_OLS,interval="confidence",se.fit=TRUE,
```

```
newdata = testing_car[, 1:5])$fit)  %>%
   mutate_if(is.numeric, round, 3)
print("Table 6. Confidence Intervals for Prediction")
head(car_cip)
```

```
[1] "Table 6. Confidence Intervals for Prediction"
```

	car_purchase_amount	age	annual_Salary	net_worth	fit	lwr	upr
1	35321.46	41.852	62812.09	238961.2	35320.87	35320.61	35321.14
2	42925.71	43.153	53798.55	638467.2	42926.31	42926.00	42926.61
3	55915.46	57.314	59729.15	560304.1	55913.03	55912.73	55913.32
4	28925.71	46.607	39814.52	326373.2	28924.16	28923.78	28924.53
5	47434.98	50.193	51752.23	629312.4	47434.92	47434.61	47435.23
6	48013.61	46.585	58139.26	630059.0	48011.63	48011.37	48011.90

```
car_pi <- testing_car   %>%
   select(car_purchase_amount, age, annual_Salary, net_worth) %>%
   cbind(predict(car_red_OLS,interval="prediction",se.fit=TRUE,
newdata = testing_car[, 1:5])$fit)  %>%
   mutate_if(is.numeric, round, 3)
print("Table 7. Prediction Intervals")
head(car_pi)
```

```
[1] "Table 7. Prediction Intervals"
```

	car_purchase_amount	age	annual_Salary	net_worth	fit	lwr	upr
1	35321.46	41.852	62812.09	238961.2	35320.87	35317.99	35323.76
2	42925.71	43.153	53798.55	638467.2	42926.31	42923.42	42929.19
3	55915.46	57.314	59729.15	560304.1	55913.03	55910.14	55915.92
4	28925.71	46.607	39814.52	326373.2	28924.16	28921.26	28927.06
5	47434.98	50.193	51752.23	629312.4	47434.92	47432.03	47437.81
6	48013.61	46.585	58139.26	630059.0	48011.63	48008.75	48014.52

Results

Looking at our model results, it is quite in line with what we were expecting. We believed that we would be able to use the attributes of a buyer to determine how much they would be willing to pay for a car. According to our models, there were attributes (3 in particular - age, annual salary, and net worth) that were statistically significant to predict the sales price. We started out by creating a full OLS regression model with all the attributes, yielding an RMSE of 1.520315. Then we wanted to see how a reduced model would perform so we used forward stepwise selection to create our OLS reduced regression model. As expected, it yielded a slightly smaller RMSE of 1.519226 as we compare in Table 5, reflecting that it is a better fit than the full OLS regression model which indicates that it will be the better model for prediction. Thus, our optimal model for predicting includes the age, annual salary, and net worth of the buyer impacting how much they spend on a car.

Discussion

Let's look at the two questions we wanted to answer and discuss what our results say about them:

1. "What attributes of a buyer can best predict how much a buyer is willing to pay for a car?": If you look at our model, it seems that age, annual salary, and net worth are the best variables to include in the prediction model, from our dataset, when estimating how much a buyer is willing to pay for a car. We came to this conclusion from both looking at:
- the exploratory data analysis (where the 3 attributes: age, annual salary, and net worth had a high positive correlation with the price spent on a car)
- our OLS full regression model (where those 3 attribute coefficients were statistically significant (p-value < 0.05)
- our OLS reduced regression (where our forward selection algorithm selected the 3 attributes listed).

1. "Based on attributes of a buyer/consumer, how much would they be willing to pay for a car?"
- This question we also answered after successfully building models with high R2's, meaning that the explanatory variables do a good job explaining the response variable compared to the null model. In the future when our team member Davis sees a customer come in and can estimate their attributes, he can use the predict function with the model (based on the customer's respective attributes) to hopefully get a good estimate of how much the customer is willing to pay for a car!

Analysing our confidence intervals for prediction (Table 6), the interpretation for row 1 confidence interval (in-sample prediction on the test set) is that with 95% confidence, the expected car_purchase amount for a person with a net-worth of 238961.2, an annual salary of 62812.09, and of 41.852 years of age is between 35320.61 and 35321.14. As well as for our prediction intervals (Table 7), the interpretation for row 1 confidence interval (in-sample prediction on the test set) is that with 95% confidence, the car_purchase amount for a person with a net-worth of 238961.2, an annual salary of 62812.09, and of 41.852 years of age is between 35317.99 and 35323.76.

As we see that age, annual salary, and net worth as the characteristics that impact car purchases, this matches in line with the study by Chandra et al. (2017) in terms of age being a factor. Our inclusion of annual salary and net worth was not considered in their study, which provides greater insight into these aspects. However, our model does not include gender as seen in their study with women having a larger pattern with age, which we suspect is due to our inclusion of variables such as salary and net worth. Additionally, our results partially disagree with the study by Rimple et al. (2015) as they state that income influenced their car purchases, and not age or gender. Although we also indicate income (in terms of annual salary and net worth) as a part of our model, age is the highest deterministic attribute from our forward selection algorithm. We suspect this is due to their small sample size in only one country which our data expands on for greater insight.

Future Steps

Given more resources and time beyond this project, we could take a few further steps to improve and expand upon our results.

1. Gather a larger dataset. Something we noted at the beginning of this project was that our dataset only consisted of 164 respondents. In the future, it would be better to collect a larger dataset for us to work with, either from another data source or through the primary collection. Large sample sizes are preferred because models created using larger datasets are usually more representative of the population, and will reduce the model's uncertainty (narrowing confidence intervals). This will make our estimates more precise.
2. Look at incorporating more variables into our model. In this dataset, we had a limited amount of variables to work with. It would be a good idea to explore other variables that may be associated with the response (for example, education) which could potentially explain the sales price of a car with more accuracy. Therefore lowering the RMSE of our model. Another limiatation/flaw to our dataset is the lack of currency indication, which may skew our results. Taking an approach with data relevant to a defined currency type would be more beneficial towards those using it as a measure. We could have also factored in differences in country, which may have also had an impact given that the cost of living differs between countries.

Our research can build upon studies that may futher look into consumer behaviour, as well as expand towards other topics including how car manufacuturers decide to choose a budget for creating cars, a target audience, and based on the attributes of their most common client groups: how they can satisfy their budget and/or expand to satisfy a larger group of audience.

Conclusion

Our study concludes that a linear model built from the age, annual salary, and net worth of a client may be a good model to predict how much someone is willing to spend on a car. Given the consumer characteristics in our model, we can predict the amount a consumer may spend on a car. Salespeople may be able to use our model to get a general idea of what prices a person is willing to pay for a car and adjust their marketing or sales strategy accordingly.

References

Chandra, A., Gulati, S., & Sallee, J. M. (2017). Who loses when prices are negotiated? an analysis of the new car market. *The Journal of Industrial Economics, 65*(2), 235-274. https://doi.org/10.1111/joie.12125.

Rimple, M., Srikant, M., Naseem, A., & Jitendra Kumar, M. (2015). A study of interaction of materialism and money attitude and its impact on car purchase. *Management & Marketing, 10*(3), 245-269. https://doi.org/10.1515/mmcks-2015-0017.

Yashpal. (n.d.). ANN - Car Sales Price Prediction, Version 1. Retrieved November 3, 2022 from https://www.kaggle.com/datasets/yashpaloswal/ann-car-sales-price-prediction.