Vick Anand - Challenge

Using multiple linear regression, design a linear model that predicts the mpg of MechaCar prototypes using a number of variables within the MechaCar mpg dataset (dependent variable). Provide a small writeup of your interpretation of the multiple linear regression results. Be sure to include the following details:

a. Which variables/coefficients provided a non-random amount of variance to the mpg values in the dataset?

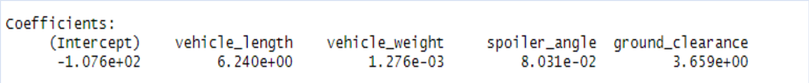
b. Is the slope of the linear model considered to be zero? Why or why not?

c. Does this linear model predict mpg of MechaCar prototypes effectively? Why or why not?

Code: Linear Multiple Regression
lm(mpg ~ vehicle_length + vehicle_weight + spoiler_angle + ground_clearance,data=MechaCar_mpg)

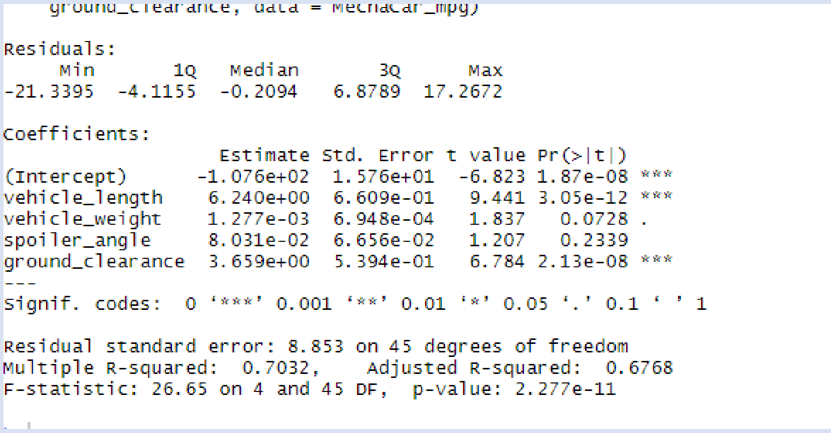
#generate summary statistics
summary(lm(mpg ~ vehicle_length + vehicle_weight + spoiler_angle + ground_clearance, data=MechaCar_mpg))

Result:1



```
Coefficients:
     (Intercept)    vehicle_length    vehicle_weight     spoiler_angle  ground_clearance
      -1.076e+02         6.240e+00         1.276e-03         8.031e-02         3.659e+00
```

Result: 2



```
    ground_clearance, data = MechaCar_mpg)

Residuals:
     Min       1Q   Median       3Q      Max 
-21.3395  -4.1155  -0.2094   6.8789  17.2672 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -1.076e+02  1.576e+01  -6.823 1.87e-08 ***
vehicle_length    6.240e+00  6.609e-01   9.441 3.05e-12 ***
vehicle_weight    1.277e-03  6.948e-04   1.837   0.0728 .  
spoiler_angle     8.031e-02  6.656e-02   1.207   0.2339    
ground_clearance  3.659e+00  5.394e-01   6.784 2.13e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.853 on 45 degrees of freedom
Multiple R-squared:  0.7032,	Adjusted R-squared:  0.6768 
F-statistic: 26.65 on 4 and 45 DF,  p-value: 2.277e-11
```

#1 Write Up

The dataset for 50 vehicles was provided with attributes such as vehicle length, vehicle weight, spoiler angle, mpg and ground clearance. We were tasked to develop linear regression model

Based on the multiple linear regression analysis, we found the following equation to predict Miles per gallon: dataset:

MPG=6.24*Vehicle_lenght+0.001*vehicle_weight+0.08031*spoiler_angle +3.66*ground_clearance - 107.6

a) The table Exhibit A below show slope or intercept of each of the values. The main drivers of predicting MPG performance is based on two main variables Vehicle Length and ground clearance.

   a. Overall: As you can see the P Value which is the probability of similar results with random sampling is quiet "low" 2.277e-11. However, the R-squared that show the strength of correlation to be 0.7 which is quite strong**. The combination of low P-Value and higher correlations results in non-random amount of variance to MPG.**

   b. At variable level: Non-random variances in two variables are dominant: **vehicle length and ground clearance that have exceptionally low P-value which are 3.05e-12 and 2.13e-08. With high 0.7 overall correlation** and low p-value it means the **model explains a lot of variation within the data and is good predictor of miles per gallon (mpg) performance metrics (best scenario)**

b) The slope of the MPG equation **is not zero** as we can see that vehicle height and ground clearance are two main variables that results in steeper slope 6.24 and 3.66 respectively.

c) Linear Model does predict MPG effectively - There is high correlation of 0.7 and lower P-value with results in non-random dataset. This means your **model explains a lot of variation within the data and is significant (best scenario).** This means equation is better predictor of mpg.

Exhibit A – Slope/Intercept and P-value for each of the independent variables

	Slope	Pr(>|t|)	Overall, p-	0.00000000002277
Intercept	(107.600)	0.00000001870		
vehicle_lenght	6.2400	0.0000000000031		
vehicle_weight	0.00128	0.0728		
spoiler_angle	0.08031	0.2339		
ground_clearance	3.65900	0.00000002130		

Suspension Coil Summary

Create a summary statistics table for the suspension coil's pounds-per-inch continuous variable with Mean, Median, Variance and Standard deviation, b) write-up of your interpretation and findings for the suspension coil summary statistics. c) Include design specifications for the MechaCar suspension coils dictate that the variance of the suspension coils must not exceed 100 pounds per inch. Does the current manufacturing data meet this design specification? Why or why not?

1) Create summary table in R:

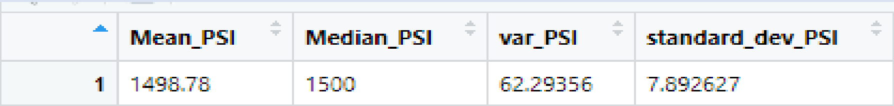
Code: Summary table with Statistics

```
coil_stats <- scoil %>% summarize(Mean_PSI=mean(PSI),Median_PSI=median(PSI),var_PSI=var(PSI),
standard_dev_PSI=sd(PSI))
```

statistics table by each of three Manufacturing Lot

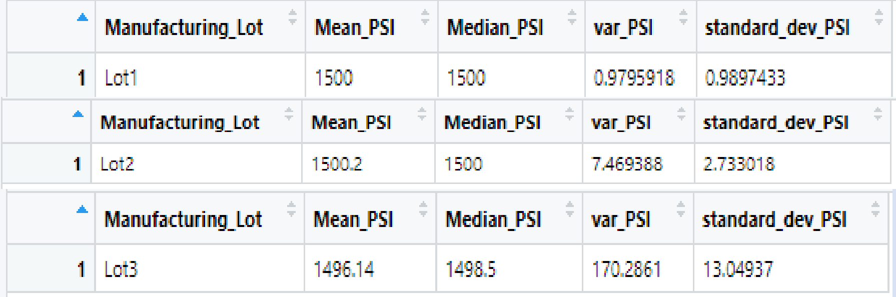
```
Lot1 <- subset(scoil,Manufacturing_Lot=="Lot1") %>% group_by(Manufacturing_Lot) %>%
summarize(Mean_PSI=mean(PSI),Median_PSI=median(PSI),var_PSI=var(PSI), standard_dev_PSI=sd(PSI))
Lot2 <- subset(scoil,Manufacturing_Lot=="Lot2") %>% group_by(Manufacturing_Lot) %>%
summarize(Mean_PSI=mean(PSI),Median_PSI=median(PSI),var_PSI=var(PSI), standard_dev_PSI=sd(PSI))
Lot3 <- subset(scoil,Manufacturing_Lot=="Lot3") %>% group_by(Manufacturing_Lot) %>%
summarize(Mean_PSI=mean(PSI),Median_PSI=median(PSI),var_PSI=var(PSI), standard_dev_PSI=sd(PSI))
```

Result:1 – Overall Summary table for the whole dataset



	Mean_PSI	Median_PSI	var_PSI	standard_dev_PSI
1	1498.78	1500	62.29356	7.892627

Result: 2 – Statics by Each Manufacturing lot to understand variation



	Manufacturing_Lot	Mean_PSI	Median_PSI	var_PSI	standard_dev_PSI
1	Lot1	1500	1500	0.9795918	0.9897433

	Manufacturing_Lot	Mean_PSI	Median_PSI	var_PSI	standard_dev_PSI
1	Lot2	1500.2	1500	7.469388	2.733018

	Manufacturing_Lot	Mean_PSI	Median_PSI	var_PSI	standard_dev_PSI
1	Lot3	1496.14	1498.5	170.2861	13.04937

Write Up

suspension coil’s pounds-per-inch (ppi) continuous variable with

1) Mean: Overall, mean of the vehicles for coils is 1,498.78. The implies that average pounds per inch of coils on each car is 1,498. Mean is calculated by adding up coils for all the cars divided by total number of vehicle ie. 50 in the dataset. We were also given the Manufacturing Lots for each of the 50 vehicles, we observed the mean for Lot 1 and 2 were 1500 almost close. Lot 3 mean was 1,496 this could mean that lot size 3 could an issue.
2) Median: Overall median for vehicle coils ppi was 1,500. This implies the value in the middle when data is sorted ascendingly for the coils. As per lot size Median value is 1500 for Lot 1 and 2 except for L3 whose median value is 1,498.5.

3) Variance: is the numerical measure of how data values are dispersed around the mean. It is useful measure of variation. Variance of a population is equal to the average squared deviation of every observation from the population mean. It is symbolized by a Greek lowercase sigma-squared (σ2). Answer c) **Overall, variation of the population is 62.29 ppi from the mean. However, when you look by lot size, the highest variation is noted in Manufacturing lot 3 which is 170.286 ppi. This is problematic as discussed below. If the design specifications state that it should not exceed 100 ppi for each vehicle, the Manufacturing Lot 3 is out of the range from the mean and can result in defects and performance issues**.

4) Standard deviation: Because variance is the average squared deviation from the mean, the units of variance are the square of the original measurements. Taking the square root of variance gives standard deviation. In R, sample standard deviation is calculated with the sd() function.

normal distribution is scaled by the standard deviation, with 68.3% of the distribution within one standard deviation of the mean, 95.4% within two standard deviations of the mean, and 99.7% within three standard. low standard deviation means that most of the numbers are close to the average. A high standard deviation means that the numbers are more spread out. Overall, standard deviation is 7.8 for 50 data population and may look normal however, when you look at lot 3, the standard deviation is high 13.04 which may be not be an ideal scenario for the when the company would like to have 1,500 coils.

In summary, Lot size 3 exceeds 100 ppi specification as variance as Lot 3 variance is 170.286 ppi. While lot 1 and 2 are in alignment with the specification. When you look at overall, the variance is only 62.29 well below 100 ppi requirement. However, you have a full batch Manufacturing lot 3 not in compliance. To avoid this problem in future, we would recommend management to conclude that current dataset not meet compliance since manufacturing lot 3 which will result in added costs and performance failures in future though overall results look favorable/in-compliance. This will avoid manufacturing lots problems. We may be able to isolate root cause of the issue by isolating problem to a particular vendor, process, or tool.

Suspension Coil T-Test

Determine if the suspension coil's pound-per-inch results are statistically different from the mean population results of 1,500 pounds per inch

Null hypothesis – Mean of Vehicle - Coil's per Pound per inch (ppi) is 1,500

H_0 = 1500 (mean represented by random chance)

Alternative hypothesis – Mean of Vehicle - Coil's per Pound per inch (ppi) is not 1,500

$H_a \neq$ 1500 (influenced by non-random events)

In order to perform t-test, this will one variable T-test. 5 conditions must be true:

1. The input data is numerical and continuous. – Vehicles count is a numerical measure and is contineous.
2. The sample data was selected randomly from its population data. – The dataset was sent to us randomly based on 50 vehicles that included the lot size.
3. The input data is considered to be normally distributed. – This is an issue with the data set. When I performed the Shapiro test, the distribution was not normal p-value < 2.2e-16. Lot 1 and 2 were both not normal distributions. However, when we did the analysis utilizing logarithmic Log 10 to normalize dataset, the results did not change much. So, I also got with Andrew (TA) who suggested to perform the analysis as is without Log10
4. The sample size is reasonably large. – the dataset had 50 vehicles
5. The variance of the input data should be very similar. – yes unit of measurement is the same ppi

Code: Summary table with Statistics

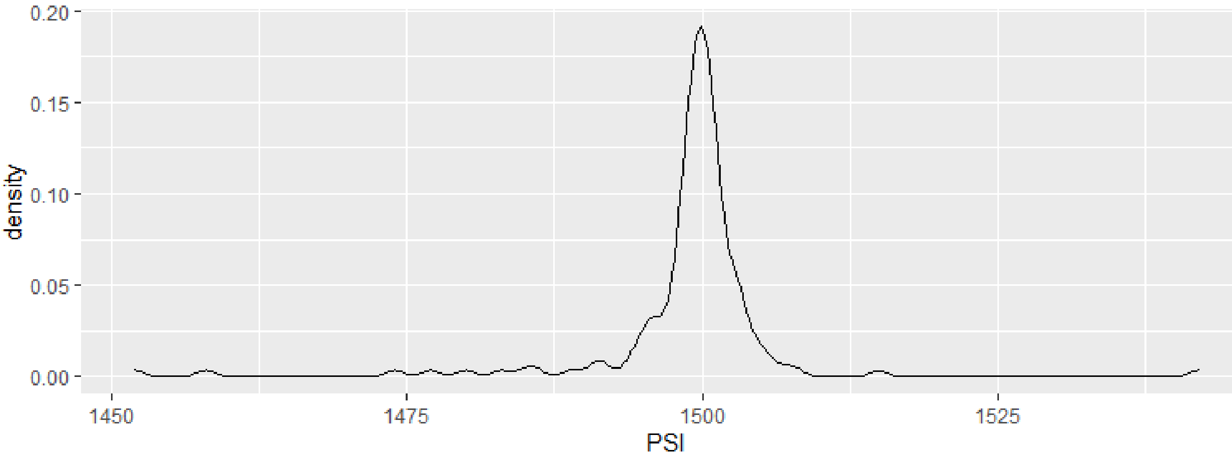
```
# 1 Step Determine Distribution - Is it Normal or not.
ggplot(scoil,aes(x=PSI))+ geom_density() # visualize distribution

# Shaparo test is seeif we have normal distribution
shapiro.test(scoil$PSI)
```

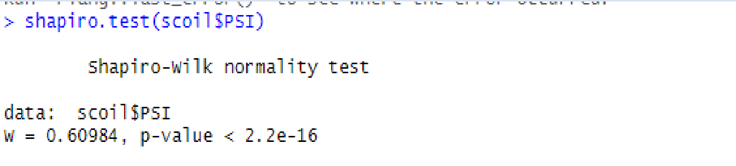
T – Test whether Mean is 1500

t.test(scoil$PSI, mu= 1500)

Result:1 – Population of Suspension Coil for Normal Distribution (Based on graph need to do Shapiro test)



Result: 2 – Shapiro Test of Population of PSI to understand Normal Distribution
Takeaway – P value very low below .05 Normal Singnifance. Therefore not a normal distribution. However, when we used Log10 to do our analysis the results were not much different. TA concluded for me to do T-test with log10.



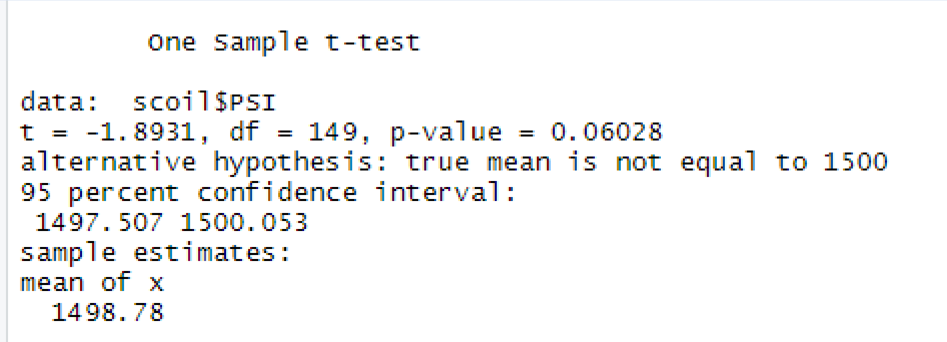
```
Run `rlang::last_error()` to see where the error occurred.
> shapiro.test(scoil$PSI)

	Shapiro-Wilk normality test

data:  scoil$PSI
W = 0.60984, p-value < 2.2e-16
```

Result: 3– T test where null to test Null Hypothesis that means is 1500
Takeaway – Based on Module when we look at P-Value is 6% which is definitely above .05 signifance level which is 5 in 100 error probable. Therefore, looking at it we may accept the mean is 1500 and **no statistical** difference between the two observed sample means.

However, we observe the mean is not equal to 1,500 it's 1,498 for the population and with 6% probability, we should keep test as "undetermined" at this time. Instead, we propose to do more testing with higher random sample set to come to this conclusion.



```
	One Sample t-test

data:  scoil$PSI
t = -1.8931, df = 149, p-value = 0.06028
alternative hypothesis: true mean is not equal to 1500
95 percent confidence interval:
 1497.507 1500.053
sample estimates:
mean of x
  1498.78
```

Design Your Own Study

Purpose: Upper management wants you to design a study that compares the performance of the MechaCar prototype vehicle to other comparable vehicles on the market to how to outperforms the competition.

- Metrics of interest to a consumer (cost, fuel efficiency, color options, etc.).
- Determine what question we would ask, what the null and alternative hypothesis would be to answer that question, and what statistical test could be used to test this hypothesis.
- Knowing what test should be used, what data should be collected? Hint: Look at the cheat sheet for required variables.

Write up

Every company has a way to outperform other company based on product differentiation. One of the aspects of product differentiation is Metrics to outperform consumer. Of course, it dependents on customer segmentation and the type of market's we want to gain market share.

Suppose our company wants to outperform in segments where car is for family:

For family-oriented customer, the following may be sample metrics, company may consider

- Extra Space – may be third row.
- Safety Setting – Rear view mirror -blind spot
- Length of Car – Aesthetics
- Miles per gallon
- Weight of the car
- Tort
- Resale value
- Average Maintenance cost per year

Above are a few important performance metrics that can capture more market share and car performance. To do just that we need to obtain dataset from our company and the customer dataset from the market sources.

The first thing to define what variables are important based on definition of our performance. We will choose variables that may be important to customer.

Isolate individual Metric that we need to approve and perform the following steps:

1) If our intent is to improve Miles per Gallon so that the vehicle provides economic value for example to customer. We need utilize dataset to identify variables that are important to improve MPG performance versus competitors

2) MPG in the examples as dependent variable and pick the variables that results in increased performance.
3) **Perform the Multi-regression test** to see which slope is steep and significant that results in improvement in metrics. To do so linear multi-regression test that will provide us with the intercept and the slope. This will also give us R-squared to show strength of correlations and the P-Value. If the P-Value is low and Correlation is high may highlight the sample is not random and is the best-case situation. With this approach, we will find which variables are the best predictors of performance MPG. Do this step for both competitor and our company and identify variable that may result in predicting the dependent variable.

4) Once you have isolated the variable, **Perform the one variable T test**, this is to compare our company mean with the other company mean. Based on T test, you could observe whether our company is exceeding the vehicle performance. In this example is MPG. The T one variable t-test is our sample dataset with **competitor dataset dependent on the following conditions:**

   a. The input data is numerical and continuous.
   b. The sample data was selected randomly from its population data.
   c. The input data is considered to be normally distributed
   d. The sample size is reasonably large.
   e. The variance of the input data should be very similar.

5) Also, we propose to perform **Anova ie. Analysis of variance testing** as another option which is used to compare the means of a continuous numerical variable across a number of groups Regardless of whichever type of ANOVA test we use, the statistical hypotheses of an ANOVA test are the same:

   H_0 : The means of all groups are equal, or $\mu_1 = \mu_2 = \ldots = \mu_n$.

   H_a : At least one of the means is different from all other groups.

6) Propose the learnings to the management and the modifications to variables required to improve performance.
7) Do trials with the proto-types, this could be done in multiple ways. One example is A/B Testing. **A/B testing** is a randomized controlled experiment that uses a control (unchanged) and experimental (changed) group to test potential changes using a success metric. A/B testing is used to test whether or not the distribution of the success metric increases in the experiment group instead of the control group; we would not want to make changes to the product that would cause a decrease in the success metric. So until proven, we don’t implement the change in production. The same goes with car software and car specification.
8) Similar to project management, if the improvement failed utilizing results from A/B testing. We take the same steps lined and revise our analysis until we obtain success.