Assignment 3: Statistics

Karla Aniela Cepeda Zapata
D00242569

Provide all the code (as a separate .py file) and report results and plots from python and interpretation of answers in a document. Load in the data “mtcars” from the csv file that is available on Moodle. Make sure to read the README attached with the data on Moodle.

Relationships

a) Briefly describe the dataset, dimensions and what type of variables there are.

Description of dataset
• The dataset is an extraction from the 1974 Motor Trend US magazine. • It is made up of fuel consumption and ten design and performance features of motor vehicles. • The dataset has 12 columns and 32 rows.

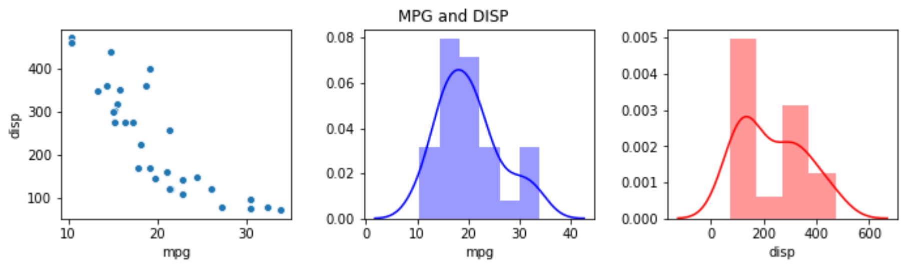
Variables				
Variable	**Metadata description**	**Additional description**	**Unit**	**Type**
Unamed :0		Probably this is the index of the original dataset. It is the brand and model of a car. I named this variable as “car” after loading, but I am not going to use it for the analysis.		Categorical Nominal
mpg	Miles per gallon	How far a car is able to travel for every gallon	Speed	Numerical Continuous
cyl	Number of cylinders	Space through which the piston travels. The number of pistons.		Numerical Discrete
disp	Displacement	Engine displacement is the combined swept volume of the pistons inside the cylinders of an engine.	Cubic inch	Numerical Continuous
hp	Gross horsepower	The power an engine produces.	Horsepower	Numerical Continuous
drat	Rear axle ratio		Ratio	Numerical Continuous
wt	Weight		Lbs	Numerical Continuous
qsec	¼ mile time	How quickly a car moves every ¼ mile	Time	Numerical Continuous
vs	Engine (0=V-shaped, 1=Straight)	Configuration for internal combustion engines.		Categorical Nominal
am	Transmission (0=automatic, 1=manual)			Categorical Nominal
gear	Number of forward gears			Numerical Discrete
carb	Number of carburetors	The device that mixes air and fuel for internal combustion.		Numerical Discrete

Additionally, we have 12 variables and 32 observations (i.e., cars).
In total, we have $12 \times 32 = 384$ data.

```
In [1513]: cars.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32 entries, 0 to 31
Data columns (total 12 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Unnamed: 0  32 non-null     object
 1   mpg         32 non-null     float64
 2   cyl         32 non-null     int64
 3   disp        32 non-null     float64
 4   hp          32 non-null     int64
 5   drat        32 non-null     float64
 6   wt          32 non-null     float64
 7   qsec        32 non-null     float64
 8   vs          32 non-null     int64
 9   am          32 non-null     int64
 10  gear        32 non-null     int64
 11  carb        32 non-null     int64
dtypes: float64(5), int64(6), object(1)
memory usage: 3.1+ KB
```

b) Investigate if any of the continuous numerical variables have a linear relationship by producing scatterplots, interpret these plots.

Plots MPG and DISP

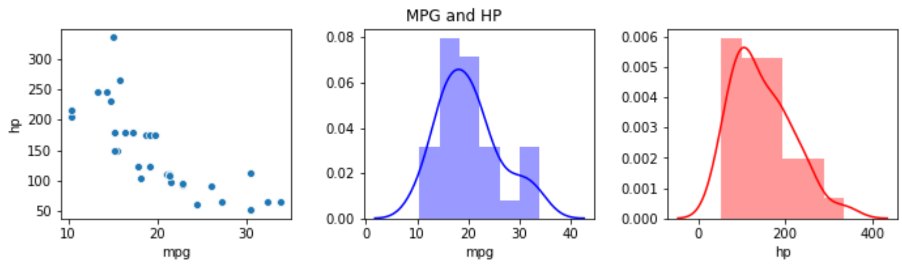



```
Statistic Summary:  mpg  and  disp
Pearson R:  (-0.8475513792624787, 9.380326537381398e-10)
Spearman S:  SpearmanrResult(correlation=-0.9088823637364655,
pvalue=6.370336284303276e-13)
Mean mpg:  20.090624999999996
Median mpg:  19.2
Mean disp:  230.72187500000004
Median disp:  196.3
```

Interpretation

- mpg has symmetric shape.
- disp is slightly skewed to the right.
- There is no linear relationship between mpg and disp. It looks exponential.

Plots MPG and HP

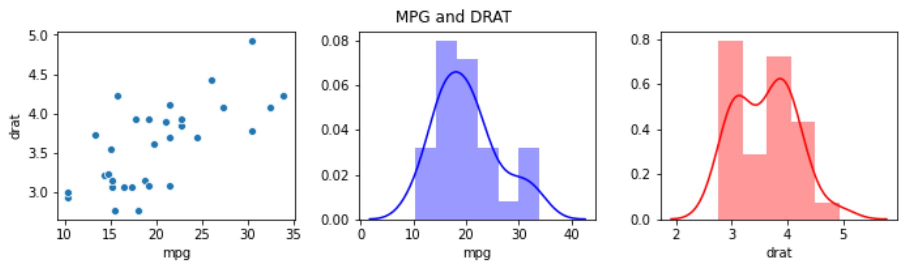



```
Statistic Summary:  mpg  and  hp
Pearson R:  (-0.7761683718265863, 1.7878352541210788e-07)
Spearman S:  SpearmanrResult(correlation=-0.8946646457499626,
pvalue=5.085969430924412e-12)
Mean mpg:  20.090624999999996
Median mpg:  19.2
Mean hp:  146.6875
Median hp:  123.0
```

Interpretation

- mpg has symmetric shape.
- hp is slightly skewed to the right.
- There is negative linear relationship between mpg and disp. (spearman: -0.8946)

Plots MPG and DRAT

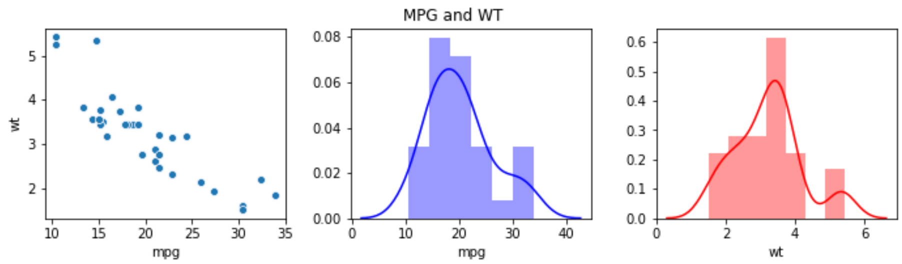



```
Statistic Summary:  mpg  and  drat
Pearson R:  (0.6811719078067493, 1.7762399287524147e-05)
Spearman S:  SpearmanrResult(correlation=0.651455464648868,
pvalue=5.381346876716847e-05)
Mean mpg:  20.090624999999996
Median mpg:  19.2
Mean drat:  3.5965625000000006
Median drat:  3.6950000000000003
```

Interpretation

- mpg has symmetric shape.
- drat has symmetric shape.
- There is a positive linear relationship between mpg and drat (pearsonr: 0.6811)

Plots MPG and WT

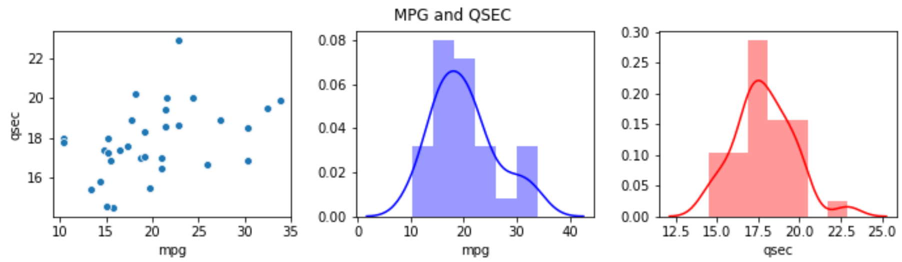



```
Statistic Summary:  mpg  and  wt
Pearson R:  (-0.8676593765172278, 1.2939587013505124e-10)
Spearman S:  SpearmanrResult(correlation=-0.8864220332702978,
pvalue=1.487594858127497e-11)
Mean mpg:  20.090624999999996
Median mpg:  19.2
Mean wt:  3.2172499999999995
Median wt:  3.325
```

Interpretation

- mpg has symmetric shape.
- drat has symmetric shape.
- There is a negative linear relationship between mpg and wt (pearsonr: -0.8676)

Plots MPG and QSEC

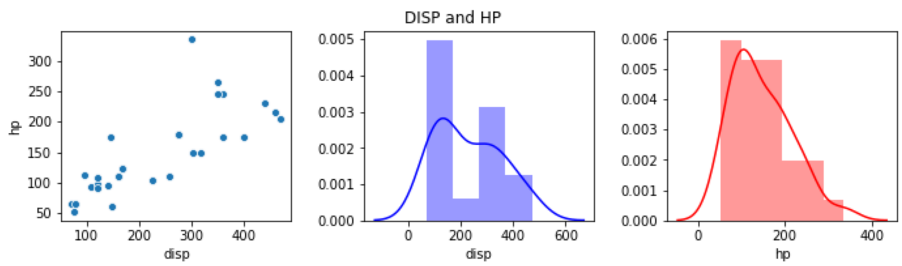



```
Statistic Summary:  mpg  and  qsec
Pearson R:  (0.4186840339217782, 0.01708198849651959)
Spearman S:  SpearmanrResult(correlation=0.4669357548683366,
pvalue=0.007055765003859273)
Mean mpg:  20.090624999999996
Median mpg:  19.2
Mean qsec:  17.848750000000003
Median qsec:  17.71
```

Interpretation

- mpg has symmetric shape.
- qsec has symmetric shape.
- There is a weak positive linear relationship between mpg and qsec (pearsonr: 0.4186)

Plots DISP and HP

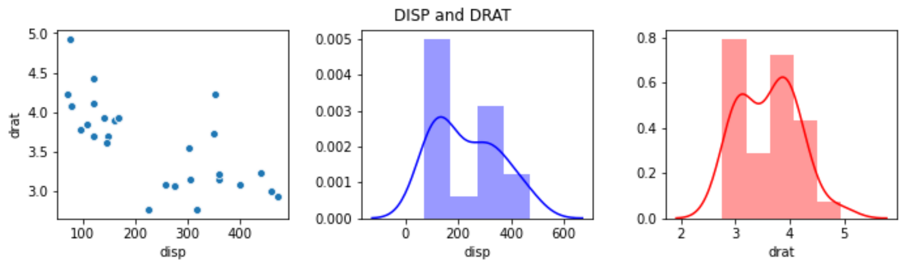



```
Statistic Summary:  disp  and  hp
Pearson R:  (0.7909485863698063, 7.142678655737436e-08)
Spearman S:  SpearmanrResult(correlation=0.8510426269929073,
pvalue=6.791337587146898e-10)
Mean disp:  230.72187500000004
Median disp:  196.3
Mean hp:  146.6875
Median hp:  123.0
```

Interpretation

- disp is slightly skewed to the right.
- hp is skewed to the right.
- There is a positive linear relationship between disp and hp (spearman: 0.8510)

Plots DISP and DRAT

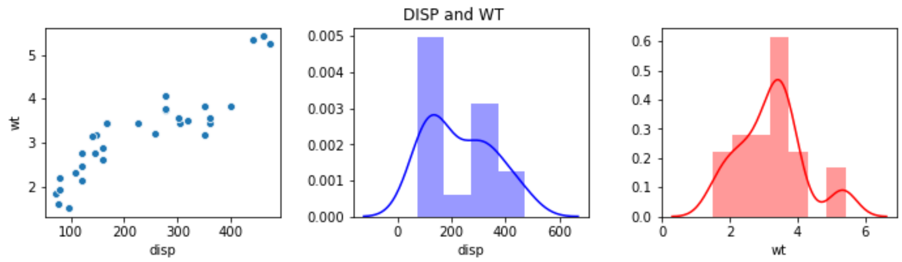



```
Statistic Summary:  disp  and  drat
Pearson R:  (-0.7102139271692702, 5.282021688157182e-06)
Spearman S:  SpearmanrResult(correlation=-0.6835920978591716,
pvalue=1.613884052358975e-05)
Mean disp:  230.72187500000004
Median disp:  196.3
Mean drat:  3.5965625000000006
Median drat:  3.6950000000000003
```

Interpretation

- disp is slightly skewed to the right.
- drat has symmetric shape.
- There is no relationship between disp and drat, it seems to have two clusters.

Plots DISP and WT

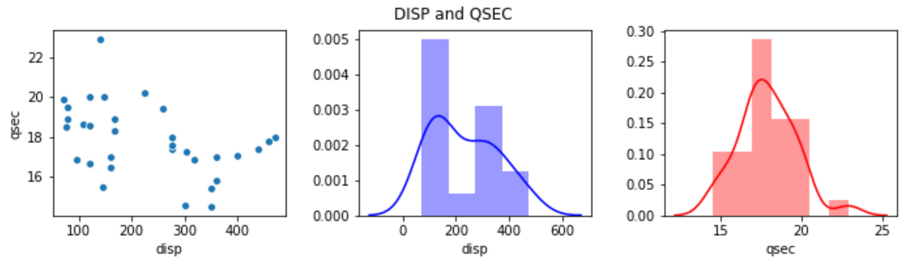



```
Statistic Summary:  disp  and  wt
Pearson R:  (0.887979922058138, 1.2223195028435877e-11)
Spearman S:  SpearmanrResult(correlation=0.8977064371299797,
pvalue=3.346361610044355e-12)
Mean disp:  230.72187500000004
Median disp:  196.3
Mean wt:  3.2172499999999995
Median wt:  3.325
```

Interpretation

- disp is slightly skewed to the right.
- wt has symmetric shape.
- There is a strong positive relationship between disp and wt (spearman: 0.8977)

Plots DISP and QSEC

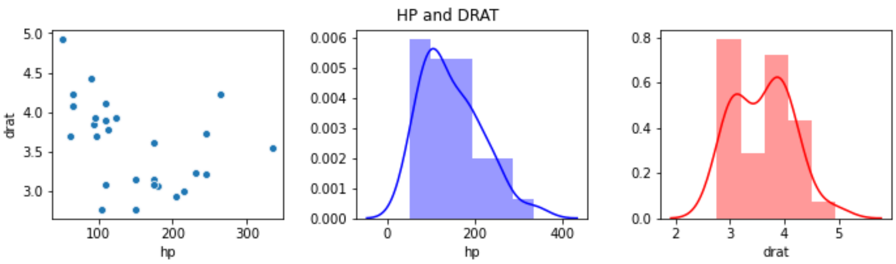



```
Statistic Summary:  disp  and  qsec
Pearson R:  (-0.4336978808110138, 0.013144036365973591)
Spearman S:  SpearmanrResult(correlation=-0.45978175980258723,
pvalue=0.008108018898408958)
Mean disp:  230.72187500000004
Median disp:  196.3
Mean qsec:  17.848750000000003
Median qsec:  17.71
```

Interpretation

- disp is slightly skewed to the right.
- qsec has symmetric shape.
- There is a weak negative relationship between disp and qsec (spearman: -0.4597)

Plots HP and DRAT

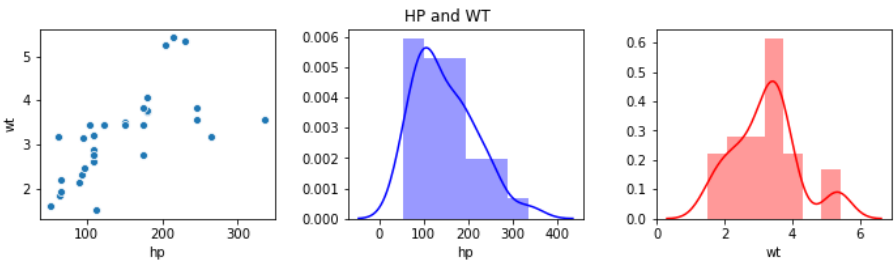



```
Statistic Summary:  hp  and  drat
Pearson R:  (-0.4487591168729195, 0.009988771893526245)
Spearman S:  SpearmanrResult(correlation=-0.5201249858108472,
pvalue=0.0022779876996084267)
Mean hp:  146.6875
Median hp:  123.0
Mean drat:  3.5965625000000006
Median drat:  3.6950000000000003
```

Interpretation

- hp is skewed to the right.
- drat has symmetric shape.
- It is not clear if there is a relationship between hp and drat. It seems to be a quadratic relationship, but stills not clear for me.

Plots HP and WT

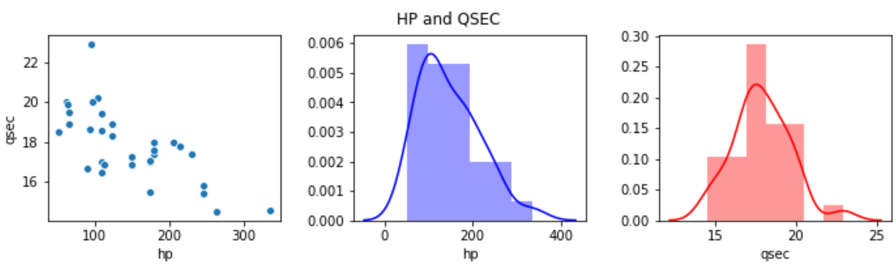



```
Statistic Summary:  hp  and  wt
Pearson R:  (0.6587478873447591, 4.145827441075327e-05)
Spearman S:  SpearmanrResult(correlation=0.7746767333913837,
pvalue=1.9537954945780572e-07)
Mean hp:  146.6875
Median hp:  123.0
Mean wt:  3.2172499999999995
Median wt:  3.325
```

Interpretation

- hp is skewed to the right.
- wt has symmetric shape.
- There is a positive linear relationship between hp and wt. (spearman: 0.7746)

Plots HP and QSEC

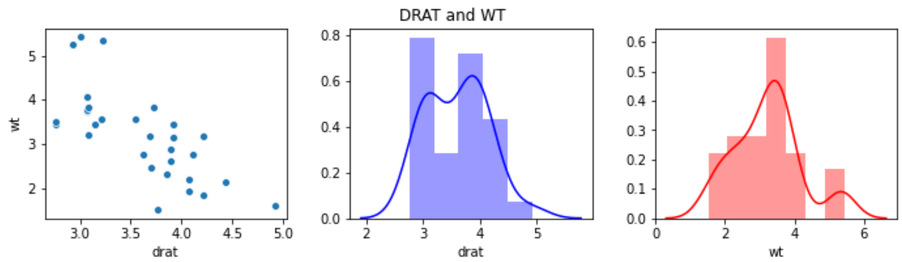



```
Statistic Summary:  hp  and  qsec
Pearson R:  (-0.7082233888619531, 5.766253455814796e-06)
Spearman S:  SpearmanrResult(correlation=-0.6666060213550505,
pvalue=3.1053442531309613e-05)
Mean hp:  146.6875
Median hp:  123.0
Mean qsec:  17.848750000000003
Median qsec:  17.71
```

Interpretation

- hp is skewed to the right.
- qsec has symmetric shape.
- There is a negative linear relationship between hp and wt. (spearman: -0.6666)

Plots DRAT and WT

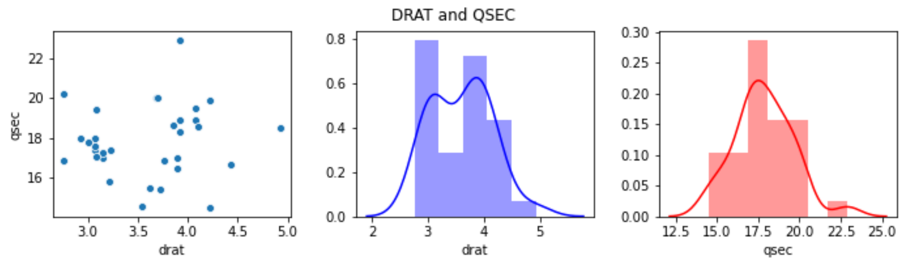



```
Statistic Summary:  drat  and  wt
Pearson R:  (-0.7124406466973718, 4.784260066132531e-06)
Spearman S:  SpearmanrResult(correlation=-0.7503904149398744,
pvalue=7.593194161573436e-07)
Mean drat:  3.5965625000000006
Median drat:  3.6950000000000003
Mean wt:  3.2172499999999995
Median wt:  3.325
```

Interpretation

- drat has symmetric shape.
- wt has symmetric shape.
- There is a negative linear relationship between drat and wt. (pearson r: -0.7124)

Plots DRAT and QSEC

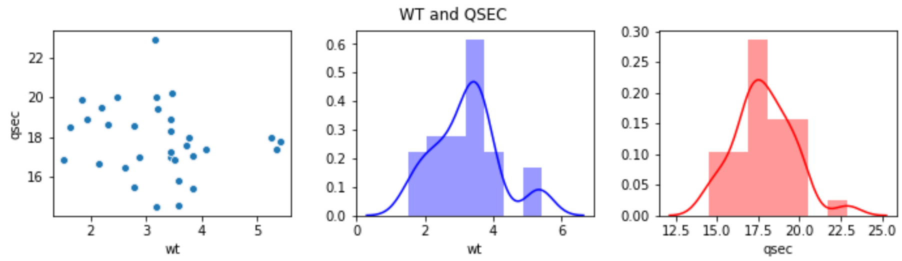



```
Statistic Summary:  drat  and  qsec
Pearson R:  (0.091204759651183, 0.6195825845550902)
Spearman S:  SpearmanrResult(correlation=0.09186863133825483,
pvalue=0.6170251388063925)
Mean drat:  3.5965625000000006
Median drat:  3.6950000000000003
Mean qsec:  17.848750000000003
Median qsec:  17.71
```

Interpretation

- drat has symmetric shape.
- qsec has symmetric shape.
- There is no relationship between drat and qsec.

Plots WT and QSEC




```
Statistic Summary:  wt  and  qsec
Pearson R:  (-0.17471587871340488, 0.33886828413491654)
Spearman S:  SpearmanrResult(correlation=-0.2254012006430882,
pvalue=0.21483883795516884)
Mean wt:  3.2172499999999995
Median wt:  3.325
Mean qsec:  17.848750000000003
Median qsec:  17.71
```

Interpretation

- wt has symmetric shape.
- qsec has symmetric shape.
- There is no relationship between drat and qsec.

c) What is the response variable and why? What is the research question of interest in this dataset?

RESPONSE VARAIBLE:

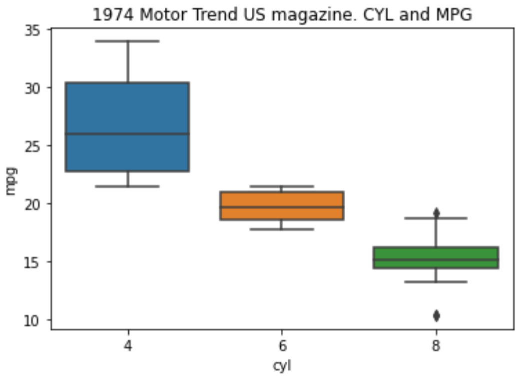
According to the scatterplots, I would pick the variables that have more linear relationships. I have three candidates: miles per gallon and weight. Both variables have three strong relationships and one weak relationship. However, the weight has more outliers, and it would be a concern for future analysis, so I would take mpg as my response variable.

RESERACH QUESTION:

In this case, we want to investigate what are the other factors that impact the value of mpg.

d) Investigate if any of the categorical /discrete variables seem to have a relationship with the response variable (selected in part b) using boxplots. Interpret these plots.

<table>
<tr><th>Boxplots</th><th>Interpretation</th></tr>
<tr><td>

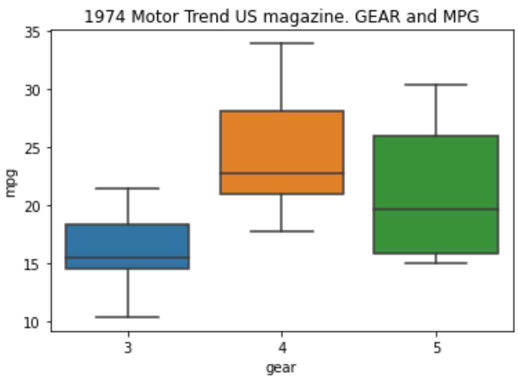


	mean	50%	std
cyl			
4	26.663636	26.0	4.509828
6	19.742857	19.7	1.453567
8	15.100000	15.2	2.560048

</td><td>

- *There is a difference between medians. In other words, there is a relationship between mpg and cyl.*
- *4-cyl is leading the group with 26.66 mpg.*
- *8-cyl is the lowest median value by 15.10 mpg.*
- *8-cyl has two outliers.*
- *There is no overlap between IQR.*
- *Spread is not the same.*
- *For group 4-cyl: roughly symmetric shape.*
- *For group 6-cyl: symmetric shape.*
- *For group 8-cyl: roughly symmetric shape.*

</td></tr>
<tr><td>

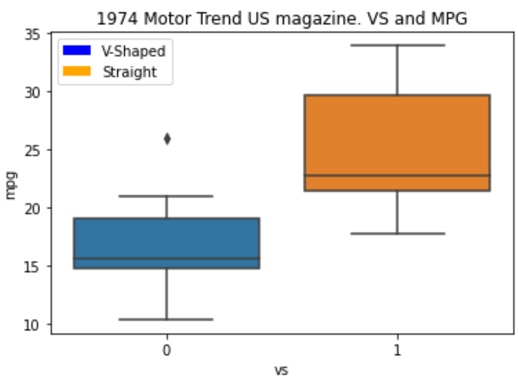


	mean	50%	std
gear			
3	16.106667	15.5	3.371618
4	24.533333	22.8	5.276764
5	21.380000	19.7	6.658979

</td><td>

- *There is a difference between medians. In other words, there is a relationship between mpg and gear.*
- *4-gear is leading the group by 24.53 mpg.*
- *3-gear is the lowest median value by 21.38 mpg.*
- *There are no outliers.*
- *IQR from 4-gear roughly overlaps IQR from 5-gear.*
- *Spread is not the same.*
- *For group 3-gear: symmetric shape.*
- *For group 4-gear: roughly symmetric shape.*
- *For group 5-gear: roughly symmetric shape.*

</td></tr>
<tr><td>

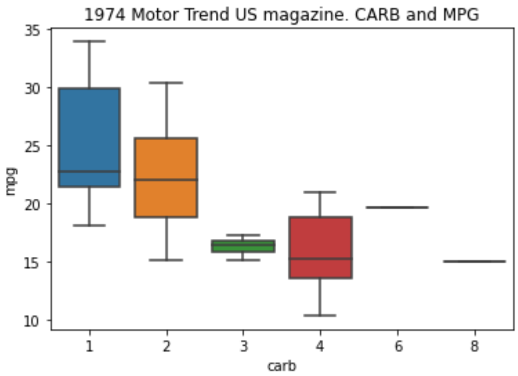


	mean	50%	std
vs			
0	16.616667	15.65	3.860699
1	24.557143	22.80	5.378978

</td><td>

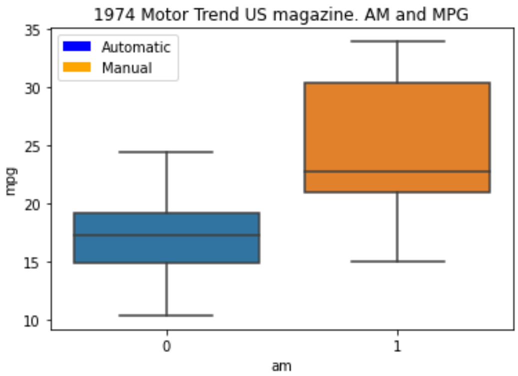
- *There is a difference between medians. In other words, there is a relationship between mpg and vs.*
- *Straight engine is leading the group by 24.55 mpg.*
- *There is one outlier in group V-Shaped.*
- *IQRs are not overlapping.*
- *Spread is roughly the same.*
- *For group V-shaped engine: roughly symmetric shape.*
- *For group straight engine: roughly symmetric shape.*

</td></tr>
</table>



carb	mean	50%	std
1	25.342857	22.80	6.001349
2	22.400000	22.10	5.472152
3	16.300000	16.40	1.053565
4	15.790000	15.25	3.911081
6	19.700000	19.70	NaN
8	15.000000	15.00	NaN

- *There is a difference between medians. In other words, there is a relationship between mpg and carbs.*
- *1-carb is leading the group by 22.80 mpg.*
- *8-cabr has the lowest median by 15 mpg.*
- *6-carb and 8-carb have just one observation.*
- *There are no outliers.*
- *2-carb is almost overlapping 1-carb.*
- *Spread is not the same.*
- *1-carb is skewed.*
- *2-carb, 3-carb and 4-carb have symmetric shape.*

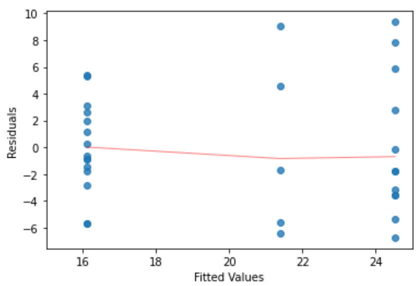
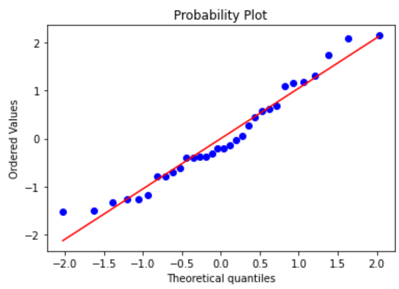


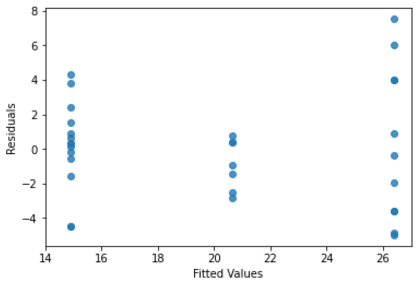
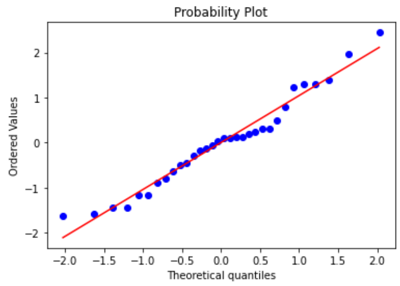
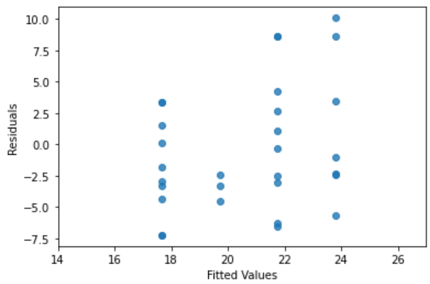
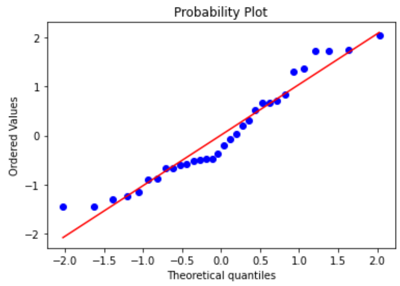
am	mean	50%	std
0	17.147368	17.3	3.833966
1	24.392308	22.8	6.166504

- *There is a difference between medians. In other words, there is a relationship between mpg and am.*
- *Manual is leading the group by 24.39 mpg.*
- *There are no outliers.*
- *IQRs are not overlapping.*
- *Spread is roughly the same.*
- *For group Automatic: symmetric shape.*
- *For group Manual: roughly symmetric shape.*

e) Using part d), choose one categorical /discrete variable with at least 3 categories to test to see if there is any difference between the means of the response variable (selected in part c) using an one-way ANOVA test. Explain your choice and hypothesis, check the assumptions and interpret the results if appropriate. Conclude your findings.

The categories that I choose to analyse with my response variable (i.e., mpg) are gear, cyl, and carbs. First of all, let's check the assumptions:

Variables	***Residual plots***	***QQ-plot***
mpg ~ gear	 *Spread is ok. We can continue.*	 *It looks normal. We can continue.*

mpg ~ cyl	 *Spread doesn't look bad. Following rule of 3:1. We can continue.*	 *Roughly all the points fall in the line. Happy with that. We can continue.*
mpg ~ carbs	 *The spread from one of the categories is too small. We cannot use rule of 3:1. Assumption violated.*	 *Roughly all the points fall in line. However, the previous assumption has been violated. We cannot continue.*

I am going to carry out a One-way ANOVA test with response variable mpg and independent variables gear and cyl. Carb variable is out of this analysis since the assumption was violated. This test is type=1 because our sample is balanced.

One-way ANOVA Test
mpg ~ gear

```
              df      sum_sq     mean_sq         F    PR(>F)
gear         1.0  259.749188  259.749188  8.995144  0.005401
Residual    30.0  866.298000   28.876600       NaN       NaN
```

Hypothsis

H0: the mpg mean are the same across each gear.
H1: at least one mpg mean across each gear is different.

Interpretation

Since P-value=0.005401< 0.05, we reject the H0. There is no evidence all mpg means across gear are equal to zero. i.e., there is a relationship between mpg and gear.

One-way ANOVA Test
mpg ~ cyl

```
              df      sum_sq     mean_sq          F        PR(>F)
cyl          1.0  817.712952  817.712952  79.561028  6.112687e-10
Residual    30.0  308.334235   10.277808        NaN           NaN
```

Hypothsis

H0: the mpg mean are the same across each cyl.
H1: at least one mpg mean across each cyl is different.

Interpretation

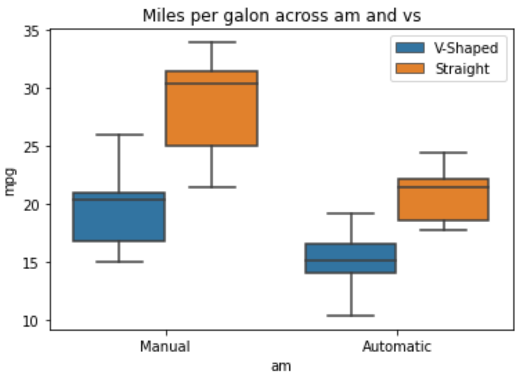
Since P-value=6.1126e-10 < 0.05, we reject the H0. There is no evidence all mpg means cross cyl are equal to zero. i.e., there is a relationship between mpg and number of cylinders.

∴ Miles per Gallon has a relationship between the forward gears and the number of cylinders. Additionally, the test couldn't be performed for the number of carburetors because one of the assumptions was violated.

f) Create a boxplot to see if there is any difference between the means of the response variable (selected in part c) across the two variables “vs” and “am”. Also, create a plot to see if there is an interaction effect for these two variables with the response variable. Interpret these plots. Test to see if any of these factors and/or interactions have a significant relationship with the response variable (selected in part c) using a two-way ANOVA test.

Boxplot	**Interaction plot**

Boxplot



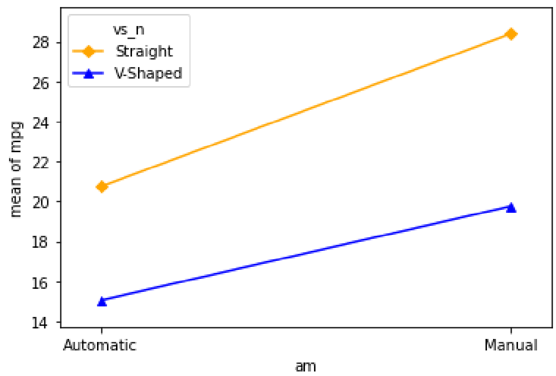
am	vs	mean	50%	std
0	0	15.050000	15.20	2.774396
	1	20.742857	21.40	2.471071
1	0	19.750000	20.35	4.008865
	1	28.371429	30.40	4.757701

Where am=0 is automatic and am=1 is manual. Where vs=0 is V-shaped engine and vs=1 is straight engine.

Interpretations:

IQRs are not overlapping. The straight engine seems to be leading where the mpg median is higher across the “am” group. It seems that "straight" engine (across any manual or automatic) influences mpg.

Interaction plot

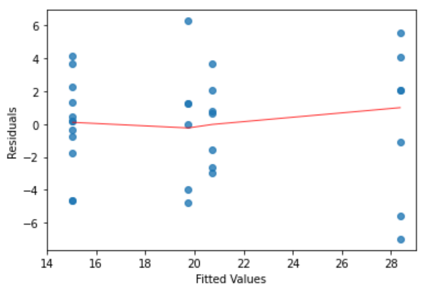
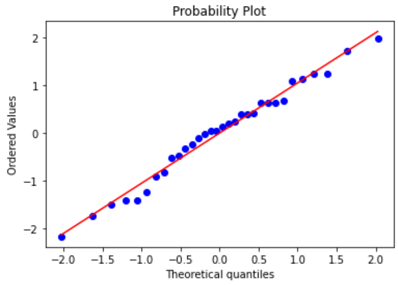


Where am=0 is automatic and am=1 is manual. Where vs=0 is V-shaped engine and vs=1 is straight engine.

Interpretation:

The means are closer for automatic comparing to manual. Lines do not seem to be parallel to each other. In this case, there might be an interaction. I am going to include it in the two-way ANOVA test.

Let’s evaluate assumptions:

Variables	***Residual plots***	***QQ-plot***
*mpg ~ C(am) * C(vs)*	 *Spread is not terrible. We can continue.*	 *It looks normal. We can continue.*

Assumptions were met, we can perform Two-way ANOVA test.

Two-way ANOVA Test
mpg ~ C(vs) * C(am)

	df	sum_sq	mean_sq	F	PR(>F)
C(am)	1.0	405.150588	405.150588	33.614841	0.000003
C(vs)	1.0	367.410647	367.410647	30.483605	0.000007
C(am):C(vs)	1.0	16.009524	16.009524	1.328290	0.258855
Residual	28.0	337.476429	12.052730	NaN	NaN

Hypothesis

In this case, we are testing three null hypothesis

Regarding groups:

1st H0: the mpg means are equal across am group (i.e., authomatic or manual.)

2nd H0: the mpg means are equal across vs group (i.e, v-shaped and straight engine.)

H1: not all mpg means are equal

Regading interaction:

H0: there is no interaction between am and vs for mpg

H1: there is interaction between am and vs for mpg

Interpretation

Since P-value=0.000003<0.05, we reject the null hypothesis. There is no evidence to suggest that mpg means across am levels are zero. Therefore, there is a relationship between mpg and am group. It is SIGNIFICANT.

Since P-value=0.000007<0.05, we reject the null hypothesis. There is no evidence to suggest that mpg means across vs group are cero. Therefore, there is a relationship between mpg and vs group. It is SIGNIFICANT

Since P-value=0.2588<0.05, we fail to reject the null hypothesis. There is not enough evidence to suggest that there is significant interaction between am and vs for mpg. Interaction between am and vs. is NOT SIGNIFICANT

Multiple Linear Regression

a) <u>Fit a suitable</u> multiple linear regression model, based on your answers from question 1, explain your choice. Comment on whether the assumptions are satisfied, interpretation of the results and the fit of the model.

For my first model, I would take all the independent variables which have a relationship with mpg, and also include categorical/discrete variables: hp, drat, wt, qsec, cyl, vs, am, gear, carb. First of all, let's check VIF values to avoid multicollinearity.

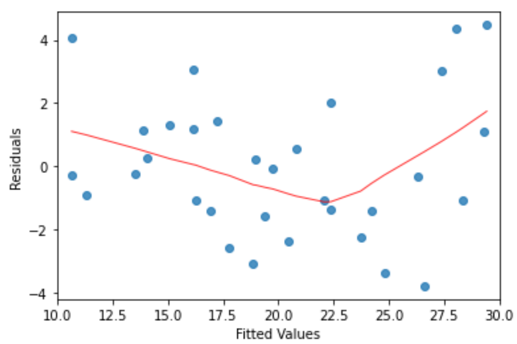
```
const    1595.773969
cyl        14.284737
hp          7.123361
qsec        6.914423
wt          6.189050
gear        5.324402
vs          4.916053
am          4.645108
carb        4.310597
drat        3.329298
```

The variable cyl has a value greater than 10. I need to remove this from the list. By running a second time, the values have changed as follows:

```
const     806.789090
wt          6.051127
hp          6.015788
qsec        5.918682
gear        4.690187
carb        4.290468
am          4.285815
vs          4.270956
drat        3.111501
```

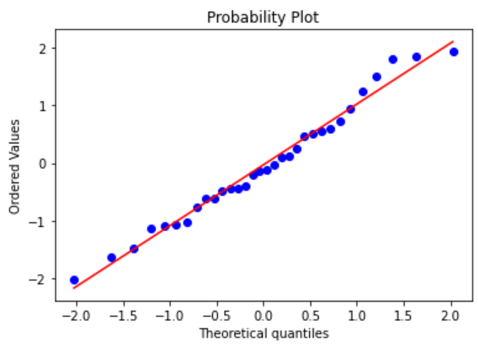
This is much better. Now, I am going to explore assumptions first:

Residual plots



It doesn't look terrible. It's ok.

QQ-plot



Not bad. Most of the points fall in the line.

Model summary

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.866
Model:                            OLS   Adj. R-squared:                  0.811
Method:                 Least Squares   F-statistic:                     15.73
Date:                Sat, 12 Dec 2020   Prob (F-statistic):           1.18e-07
Time:                        14:31:57   Log-Likelihood:                -70.274
No. Observations:                  32   AIC:                             160.5
Df Residuals:                      22   BIC:                             175.2
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         12.5505     18.526      0.677      0.505     -25.870      50.971
hp            -0.0129      0.018     -0.706      0.488      -0.051       0.025
drat           0.9286      1.608      0.578      0.569      -2.406       4.263
wt            -2.6269      1.198     -2.193      0.039      -5.111      -0.142
qsec           0.6652      0.693      0.959      0.348      -0.773       2.103
cyl            0.0963      0.997      0.097      0.924      -1.972       2.164
vs             0.1604      2.073      0.077      0.939      -4.138       4.459
am             2.4788      2.035      1.218      0.236      -1.742       6.699
gear           0.7430      1.474      0.504      0.619      -2.313       3.799
carb          -0.6169      0.606     -1.018      0.320      -1.873       0.639
==============================================================================
Omnibus:                        1.268   Durbin-Watson:                   1.875
Prob(Omnibus):                  0.531   Jarque-Bera (JB):                1.227
Skew:                           0.405   Prob(JB):                        0.541
Kurtosis:                       2.486   Cond. No.                     6.51e+03
==============================================================================
```

Residual standard error

```
In [50]: np.sqrt(model_mlr.scale)
Out[50]: 2.566296947883662
```

Fit model
$mpg = 12.5505 - 0.0129hp + 0.9286drat - 2.6259wt + 0.6652qsec + 0.0963cyl + 0.1604vs + 2.4788am + 0.7430gear - 0.6169carb + e$
Interpretation of Fit model
• 12.5505 is the average mpg when all independent variables are equal to zero. It does not make much sense since cyl, gear, and carb most have a value greater than zero. • For every unit of power increase in horsepower, the average mile per gallon decreases 0.0129, keeping the rest of the independent variables the same. • For every decimal increase in rear axle ratio, the average mpg increases 0.9286, keeping the rest of the independent variables the same. • For every lbs increase in weight, the average mile per gallon decreases 2.6259, keeping the rest of the independent variables the same. • For every second increase in ¼ mile time, the average mpg increases 0.6652, keeping the rest of the independent variables the same. • Going from 6, 4 and 8 cylinders, the average mile per gallon increases 0.0963, keeping the rest of the independent variables the same. • A car that has a straight engine is associated with how far a car is able to travel, increasing 0.963 the average mpg, keeping the rest of the independent variables the same. • A car that has a manual transmission is associated with how far a car is able to travel, increasing 2.4788 the average mpg, keeping the rest of the independent variables the same. • Going from 4, 3, and 5 forward gear, the average mile per gallon increases 0.7430, keeping the rest of the independent variables the same. • Going from 1 to 8 carburets, the average mile per gallon decreases 0.6169, keeping the rest of the independent variables the same.
Significant independent variables
Hypothesis: $H_0: \beta_i = 0$ $H_1: \beta_i \neq 0$ Looking at the regression coefficients result: • $\hat{\beta}_1$, horsepower, has a p-value = 0.488 > 0.5. Therefore, we fail to reject the null hypothesis and conclude this independent variable is not significant. • $\hat{\beta}_2$, rear axle ratio, has a p-value = 0.569 > 0.5. Therefore, we fail to reject the null hypothesis and conclude this independent variable is not significant. • $\hat{\beta}_3$, weight, has a p-value = 0.039 < 0.5. Therefore, we reject the null hypothesis and conclude this independent variable is significant. • $\hat{\beta}_4$, ¼ mile time, has a p-value = 0.348 > 0.5. Therefore, we fail to reject the null hypothesis and conclude this independent variable is not significant. • $\hat{\beta}_5$, number of cylinders, has a p-value = 0.924 > 0.5. Therefore, we fail to reject the null hypothesis and conclude this independent variable is not significant. • $\hat{\beta}_6$, V-shaped or straight engine, has a p-value = 0.939 > 0.5. Therefore, we fail to reject the null hypothesis and conclude this independent variable is not significant. • $\hat{\beta}_7$, automatic or manual, has a p-value = 0.236 > 0.5. Therefore, we fail to reject the null hypothesis and conclude this independent variable is not significant. • $\hat{\beta}_8$, number of forward gears, has a p-value = 0.619 > 0.5. Therefore, we fail to reject the null hypothesis and conclude this independent variable is not significant. • $\hat{\beta}_9$, number of carburetors, has a p-value = 0.320 > 0.5. Therefore, we fail to reject the null hypothesis and conclude this independent variable is not significant.

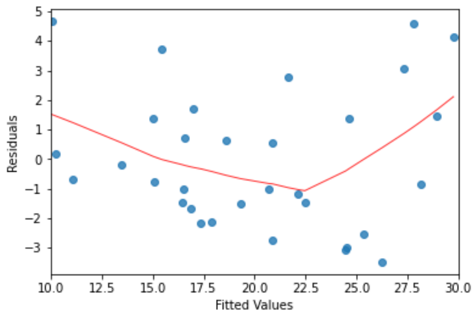
b) Are there any variables you would like to remove/add from/to the model and why? Re-run the multiple linear regression model. (You can do this more than once, in a background stepwise elimination if you think appropriate, to find the most suitable model). Compare the fit of this model to the model in part a. Perform an F-test to compare the two models, stating your hypothesis and the conclusion of this test.

Definitely, there is so much to do with the model, since I got many independent variables which are not significant to the model. I am going to start by removing the highest p-value. In the following table, I recorded the R-square, F-statistic, Standard Error, AIB and BIC:

Model	*Adj. R-square*	*F-Statistic P-value*	*RSE*	*AIB*	*BIC*	*Comment*
$mpg = 12.5505 - 0.0129hp + 0.9286drat - 2.6259wt + 0.6652qsec + 0.0963cyl + 0.1604vs + 2.4788am + 0.7430gear - 0.6169carb + e$	***0.811***	***1.18e-07***	***2.56629***	***160.5***	***175.2***	***Original model I tried in previous question. Let's remove cyl variable which has the highest P-value.***
$mpg = 13.8081 - 0.0123hp + 0.8889drat - 2.6097wt + 0.6398qsec + 0.0879vs + 2.4242am + 0.6939gear - 0.6129carb + e$	***0.865***	***2.62e-08***	***2.56629***	***158.6***	***171.8***	***There are still variables with high p-values. I'll remove vs variable which has the highest P-value***
$mpg = 13.4921 - 0.0121hp + 0.8976drat - 2.6277wt + 0.6585qsec + 2.4135am + 0.7055gear - 0.6146carb + e$	***0.865***	***5.31e-09***	***2.51238***	***156.6***	***168.3***	***There are still variables with high p-values. I'll remove gear variable which has the highest P-value***
$mpg = 14.5511 - 0.0117hp + 1.0728drat - 2.8290wt + 0.7071qsec + 2.8586am - 0.4544carb + e$	***0.864***	***1.12e-09***	***2.47663***	***155***	***165.2***	***There are still variables with high p-values. I'll remove hp variable which has the highest P-value***
$mpg = 9.9243 + 1.2071drat - 3.1108wt + 0.9145qsec + 2.9639am - 0.6023carb + e$	***0.861***	***2.42e-10***	***2.45538***	***153.7***	***162.5***	***There are still variables with high p-values. I'll remove drat variable which has the highest P-value***
$mpg = 12.8972 - 3.4343wt + 1.0191qsec + 3.5114am - 0.4886carb + e$	***0.857***	***5.06e-11***	***2.44381***	***152.6***	***159.9***	***There are still variables with high p-values. I'll remove carb variable which has the highest P-value***
$mpg = 9.6178 - 3.9165wt + 1.2259qsec + 2.9358am + e$	***0.834***	***1.21e-11***	***2.45884***	***152.1***	***158.0***	***All variables are significant. Let's see assumptions.***

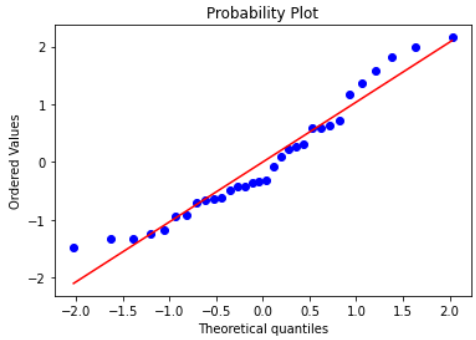
Reduced model

Residual plots



It doesn't look terrible. It's ok.

QQ-plot



Not bad. Most of the points fall in the line.

Model summary

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.850
Model:                            OLS   Adj. R-squared:                  0.834
Method:                 Least Squares   F-statistic:                     52.75
Date:                Sat, 12 Dec 2020   Prob (F-statistic):           1.21e-11
Time:                        17:18:12   Log-Likelihood:                -72.060
No. Observations:                  32   AIC:                             152.1
Df Residuals:                      28   BIC:                             158.0
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          9.6178      6.960      1.382      0.178      -4.638      23.874
wt            -3.9165      0.711     -5.507      0.000      -5.373      -2.460
qsec           1.2259      0.289      4.247      0.000       0.635       1.817
am             2.9358      1.411      2.081      0.047       0.046       5.826
==============================================================================
Omnibus:                        2.574   Durbin-Watson:                   1.714
Prob(Omnibus):                  0.276   Jarque-Bera (JB):                2.213
Skew:                           0.540   Prob(JB):                        0.331
Kurtosis:                       2.297   Cond. No.                         296.
==============================================================================
```

Residual standard error

```
In [108]: np.sqrt(model_mlr.scale)
Out[108]: 2.4588464886845087
```

Fit model

$$mpg = 9.6178 - 3.9165wt + 1.2259qsec + 2.9358am + e$$

Interpretation of Fit model

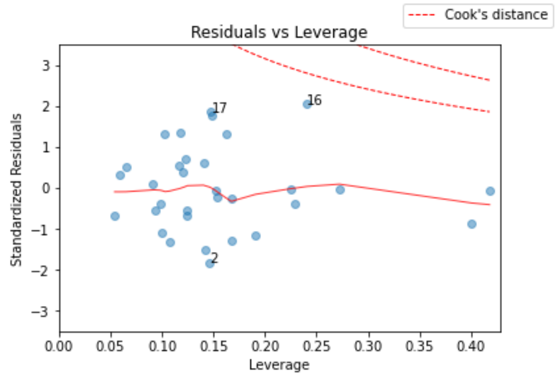
- 9.6178 is the average mpg when all independent variables are equal to zero. It does not make much sense since weight cannot be zero.
- For every lbs increase in weight, the average mile per gallon decreases 3.9165, keeping the rest of the independent variables the same.
- For every second increase in ¼ mile time, the average mpg increases 1.2259, keeping the rest of the independent variables the same.
- A car which has a manual transmission is associated with how far a car is able to travel, increasing 2.9358 the average mpg, keeping the rest of the independent variables the same.

Significant independent variables

Hypothesis:
$H_0: \beta_i = 0$
$H_1: \beta_i \neq 0$
Looking at the regression coefficients result:

- $\hat{\beta}_1$, weight, has a p-value = 0.000 < 0. Therefore, we reject the null hypothesis and conclude this independent variable is significant.
- $\hat{\beta}_2$, ¼ mile time, has a p-value = 0.000 < 0. Therefore, we reject the null hypothesis and conclude this independent variable is significant.
- $\hat{\beta}_3$, automatic or manual, has a p-value = 0.047 < 0.5. Therefore, we reject the null hypothesis and conclude this independent variable is significant.

Let's have a quick look at Cook's Distance line and check if the reduced model has influence points (i.e., leverage and outliers). As shown in the plot below, there are no points above the Cook's distance line.



Additionally, by looking at following:

Model	***Adj. R-square***	***F-Statistic P-value***	***Residual Standard Error***	***AIB***	***BIC***
$mpg = 12.5505 - 0.0129hp + 0.9286drat - 2.6259wt + 0.6652qsec + 0.0963cyl + 0.1604vs + 2.4788am + 0.7430gear - 0.6169carb + e$	***0.811***	***1.18e-07***	***2.56629***	***160.5***	***175.2***
$mpg = 9.6178 - 3.9165wt + 1.2259qsec + 2.9358am + e$	***0.834***	***1.21e-11***	***2.45884***	***152.1***	***158.0***

- ***Variability is better explained by the reduced model with 83.4%.***
- ***The Residual Standard Error value in the reduced model is smaller.***
- ***The AIB and the BIC values are also smaller in the reduced model comparing with the full model.***

Now, let's perform the F-test:

<table>
<tr><th>F-test</th></tr>
<tr><td>

```
In [121]: print(anovaResults)
   df_resid         ssr  df_diff   ss_diff         F    Pr(>F)
0      27.0  161.249986      0.0       NaN       NaN       NaN
1      23.0  151.475241      4.0  9.774745  0.371049  0.826787
```

</td></tr>
<tr><td>Hypothesis:
$H_0: Reduced\ model\ is\ preferred$
$H_1: Reduced\ model\ is\ not\ preferred$
Conclusions:
• P-value = 0.826787 > 0.05. Therefore, we fail to reject the null hypothesis and conclude there is no evidence to suggest full model is more appropriate.</td></tr>
</table>

∴ Looking at the AIB, BIC, Adj. R-Square, Residual Standard Error and the F-test, the reduced model is preferred.

c) Conclude your overall results. Use your preferred model to predict the fitted values for your response variable and calculate the residual term if you are given the following data:

mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	Carb
27	4	143	89	3.85	2.95	27.8	1	1	4	2

In conclusion, the reduced model is preferred since when I performed the F-test, it has shown that there is no evidence to suggest the full model is more appropriate. Additionally, AIB and BIC from the reduced model have shown a smaller value than the full model. Moreover, variability is better explained by the reduced model comparing with the full model with 84%. Finally, the Residual Standard Error value is smaller in the reduced model.

Reduce model to predict fitted values:

$$\boldsymbol{mpg = 9.6178 - 3.9165wt + 1.2259qsec + 2.9358am + e}$$

To calculate fitter value from data given:

$$mpg = 9.617 - 3.9156 \times 2.95 + 1.2259 \times 27.8 + 2.9358 \times 1$$
$$mpg = 35.0818$$

To calculate residual term from data given:

$$Residual(error) = e = y - \hat{y}$$
$$e = 27 - 35.0818$$
$$e = -8.079944$$

From sample:

	mpg	**wt**	**qsec**	**am**	$\hat{y}$	$e = y - \hat{y}$
Mazda RX4	21	2.62	16.46	1	22.470684	*-1.470684*
Mazda RX4 Wag	21	2.875	17.02	1	22.1584805	*-1.1584805*
Datsun 710	22.8	2.32	18.61	1	26.281319	*-3.481319*
Hornet 4 Drive	21.4	3.215	19.44	0	20.8577485	*0.5422515*
Hornet Sportabout	18.7	3.44	17.02	0	17.009858	*1.690142*
Valiant	18.1	3.46	20.22	0	20.854408	*-2.754408*
Duster 360	14.3	3.57	15.84	0	15.054151	*-0.754151*
Merc 240D	24.4	3.19	20	0	21.642165	*2.757835*
Merc 230	22.8	3.15	22.9	0	25.353935	*-2.553935*
Merc 280	19.2	3.44	18.3	0	18.57901	*0.62099*
Merc 280C	17.8	3.44	18.9	0	19.31455	*-1.51455*
Merc 450SE	16.4	4.07	17.4	0	15.008305	*1.391695*
Merc 450SL	17.3	3.73	17.6	0	16.585095	*0.714905*
Merc 450SLC	15.2	3.78	18	0	16.87963	*-1.67963*
Cadillac Fleetwood	10.4	5.25	17.98	0	11.097857	*-0.697857*
Lincoln Continental	10.4	5.424	17.82	0	10.220242	*0.179758*
Chrysler Imperial	14.7	5.345	17.42	0	10.0392855	*4.6607145*
Fiat 128	32.4	2.2	19.47	1	27.805573	*4.594427*
Honda Civic	30.4	1.615	18.52	1	28.9321205	*1.4678795*
Toyota Corolla	33.9	1.835	19.9	1	29.7622325	*4.1377675*
Toyota Corona	21.5	2.465	20.01	0	24.4938865	*-2.9938865*
Dodge Challenger	15.5	3.52	16.87	0	16.512653	*-1.012653*
AMC Javelin	15.2	3.435	17.3	0	17.3726925	*-2.1726925*
Camaro Z28	13.3	3.84	15.41	0	13.469559	*-0.169559*
Pontiac Firebird	19.2	3.845	17.05	0	15.4604525	*3.7395475*
Fiat X1-9	27.3	1.935	18.9	1	28.1446825	*-0.8446825*
Porsche 914-2	26	2.14	16.7	1	24.64482	*1.35518*
Lotus Europa	30.4	1.513	16.9	1	27.3456455	*3.0543545*
Ford Pantera L	15.8	3.17	14.5	1	17.913845	*-2.113845*
Ferrari Dino	19.7	2.77	15.5	1	20.706345	*-1.006345*
Maserati Bora	15	3.57	14.6	1	16.469835	*-1.469835*
Volvo 142E	21.4	2.78	18.6	1	24.46747	*-3.06747*