Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer

Seasons:

- In Summer the count is very low.
- In Spring and Fall the counts are better than Summer
- In Winter the counts are better than Summer but not as good as Spring and Fall

Year:

- The counts of 2019 are better as compared to 2018

Month:

- The count variable is the least in Jan and increases in the middle months (Mar-Oct) and then reduces slightly in Nov and Dec

Holiday:

- The mean count is higher in non-Holiday as compared to holiday

Weekday and Workday:

- The mean count is almost the same in weekday and workday

Weather:

- Count is maximum when weather is good and its minimum weather is bad.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

The drop_first= True attribute will remove the first dummy variable. This will help us represent the entire data with one less variable. For example, if you have a month column in your data, having values from 1 to 12 for each Jan to Dec month respectively. So instead of creating 12 dummy variables for month categorical column we can just create n-1 i.e 11 columns by dropping the first dummy variable.

For example, if the Jan column is dropped, all the remaining 11 columns will have 0 as value which means that it represents Jan. Below is the list of values for each month after creating the dummy variable and dropping the first i.e Jan from the dataframe.

Jan- 00000000000
Feb- 10000000000
Mar- 01000000000
Apr- 00100000000
May- 00010000000
Jun- 00001000000

Jul- 00000100000
Aug- 00000010000
Sep- 00000001000
Oct- 00000000100
Nov- 00000000010
Dec- 00000000001

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Temperature has the highest correlation (0.63) among all the variables.

Correlation matrix

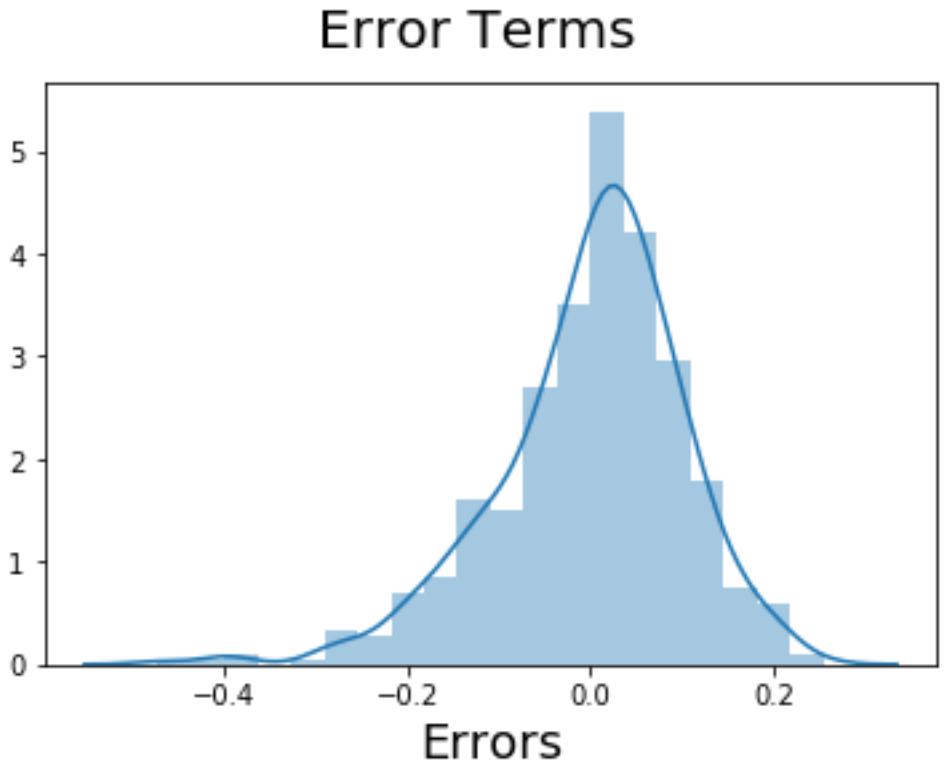
	season	year	month	holiday	weekday	workingday	weather	temperature	atemp	humidity	windspeed	count
season	1	-0.018	0.83	-0.011	-0.0031	0.012	0.019	0.33	0.34	0.21	-0.23	0.41
year	-0.0018	1	-0.018	0.008	-0.0055	-0.002	-0.049	0.048	0.046	-0.11	-0.012	0.57
month	0.83	-0.018	1	0.019	0.0095	-0.0059	0.044	0.22	0.23	0.22	-0.21	0.28
holiday	-0.011	0.008	0.019	1	-0.1	-0.25	-0.035	-0.029	-0.033	-0.016	0.0063	-0.068
weekday	-0.0031	-0.0055	0.0095	-0.1	1	0.036	0.031	-0.00017	-0.0075	-0.052	0.014	0.067
workingday	0.012	-0.002	-0.0059	-0.25	0.036	1	0.061	0.053	0.052	0.024	-0.019	0.061
weather	0.019	-0.049	0.044	-0.035	0.031	0.061	1	-0.12	-0.12	0.59	0.04	-0.3
temperature	0.33	0.048	0.22	-0.029	-0.00017	0.053	-0.12	1	0.99	0.13	-0.16	0.63
atemp	0.34	0.046	0.23	-0.033	-0.0075	0.052	-0.12	0.99	1	0.14	-0.18	0.63
humidity	0.21	-0.11	0.22	-0.016	-0.052	0.024	0.59	0.13	0.14	1	-0.25	-0.1
windspeed	-0.23	-0.012	-0.21	0.0063	0.014	-0.019	0.04	-0.16	-0.18	-0.25	1	-0.23

Correlation matrix

	season	year	month	holiday	weekday	workingday	weather	temperature	atemp	humidity	windspeed	count
count	0.41	0.57	0.28	-0.068	0.067	0.061	-0.3	0.63	0.63	-0.1	-0.23	1

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

We verified that the Error terms are Normally distributed. The graph below shows that the errors terms are normally distributed.



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

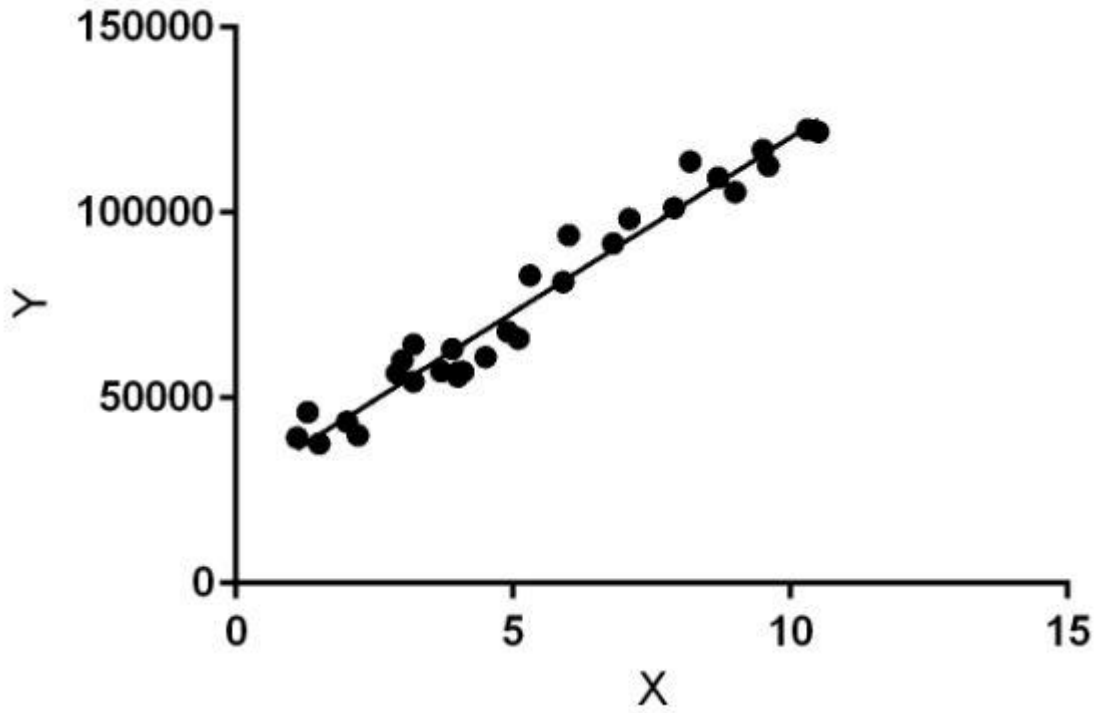
Temperature, Year_2019 and Summer are the 3 most significantly explain the demand of the shared bikes

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

   Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between

dependent and independent variables, they are considering and the number of independent variables being used.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.
In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Hypothesis function for Linear Regression :

$$y = \theta_1 + \theta_2.x$$

While training the model we are given :
x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)
When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ_1 and θ_2 values.
θ_1: intercept
θ_2: coefficient of x
Once we find the best θ_1 and θ_2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (*x*,*y*) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to

counter the impression among statisticians that "numerical calculations are exact, but graphs are rough.

Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x,y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize **COMPLETELY,** when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

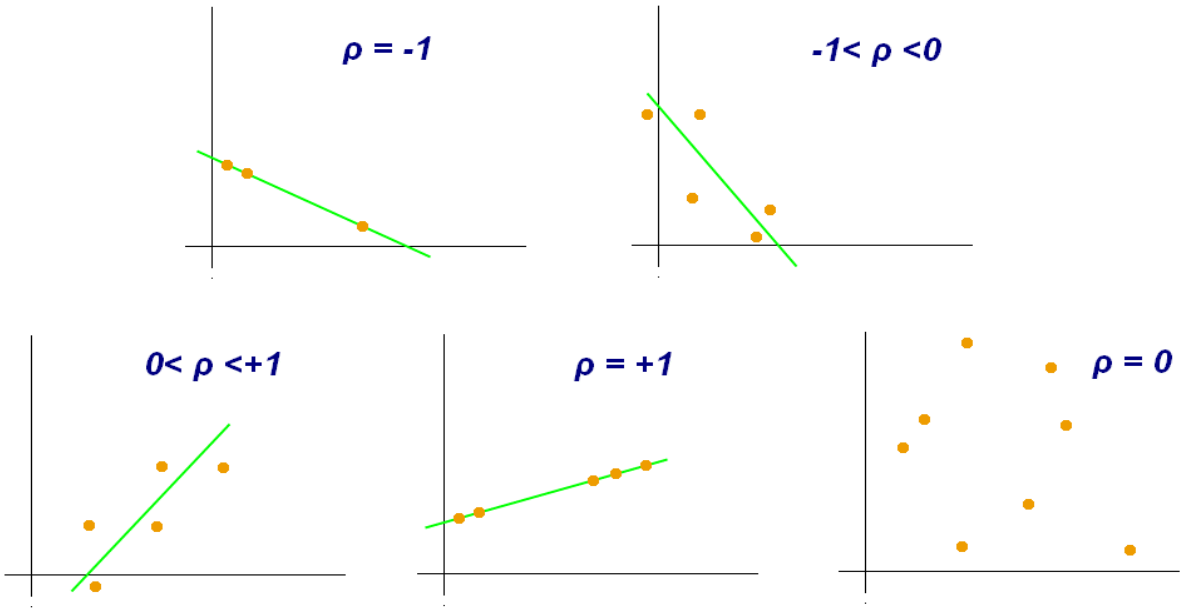
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

Quartet's Summary Stats

3. What is Pearson's R? (3 marks)

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's *r*, the Pearson product-moment correlation coefficient (PPMCC) or the bivariate correlation, is a measure of the linear correlation between two variables *X* and *Y*. According to the Cauchy–Schwarz inequality it has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. It is widely used in the sciences.

Examples of scatter diagrams with different values of correlation coefficient (ρ)



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

Techniques to perform Feature Scaling

Consider the two most important ones:

- **Min-Max Normalization:** This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

- **Standardization:** It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

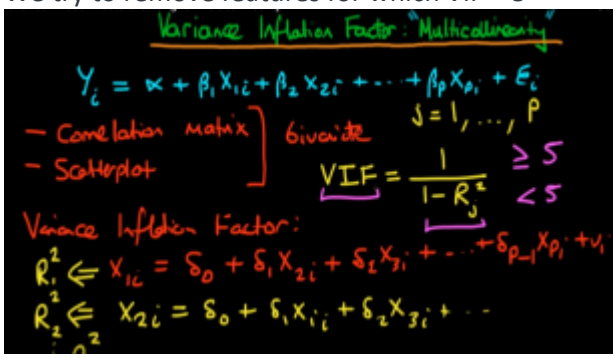
VIF = Variance Inflation Factor

In linear regression collinearity can make coefficient unstable

There will not be any issue in prediction accuracy but coefficients would be less reliable and p-value would be more
Correlation coefficients help us detect correlation between pairs but not the multiple correlation x1 = 2*x3 + 4*x7
PCA is one thing, we don't want to transform variable to keep interpret-ability intact
We want some way to reduce dimensions
In VIF, each feature is regression against all other features. If R2 is more which means this feature is correlated with other features.

VIF = 1 / (1 – R2)

When R2 reaches 1, VIF reaches infinity
We try to remove features for which VIF > 5



Example at [1] shows the use of VIF to reduce no of features.
Once we identify high VIF for features we need to reduce it
We can do it by eliminating some features
How to identify which feature to remove?
Check the correlated features for feature having high VIF
In the example at [1] weighted and BSA were correlated
Practically it is easy to measure weight so we kept it
So such decision depends on the practical implication

There can be the case that one feature is correlated with many others and we might want to remove it

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-12.87	2.56	-5.03	0.000	
Age	0.7033	0.0496	14.18	0.000	1.76
Weight	0.9699	0.0631	15.37	0.000	8.42
BSA	3.78	1.58	2.39	0.033	5.33
Dur	0.0684	0.0484	1.41	0.182	1.24
Pulse	-0.0845	0.0516	-1.64	0.126	4.41
Stress	0.00557	0.00341	1.63	0.126	1.83

Correlation: BP, Age, Weight, BSA, Dur, Pulse

	BP	Age	Weight	BSA	Dur	Pulse
Age	0.659					
Weight	0.950	0.407				
BSA	0.866	0.378	0.875			
Dur	0.293	0.344	0.201	0.131		
Pulse	0.721	0.619	0.659	0.465	0.402	
Stress	0.164	0.368	0.034	0.018	0.312	0.506

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(3 marks)

he Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

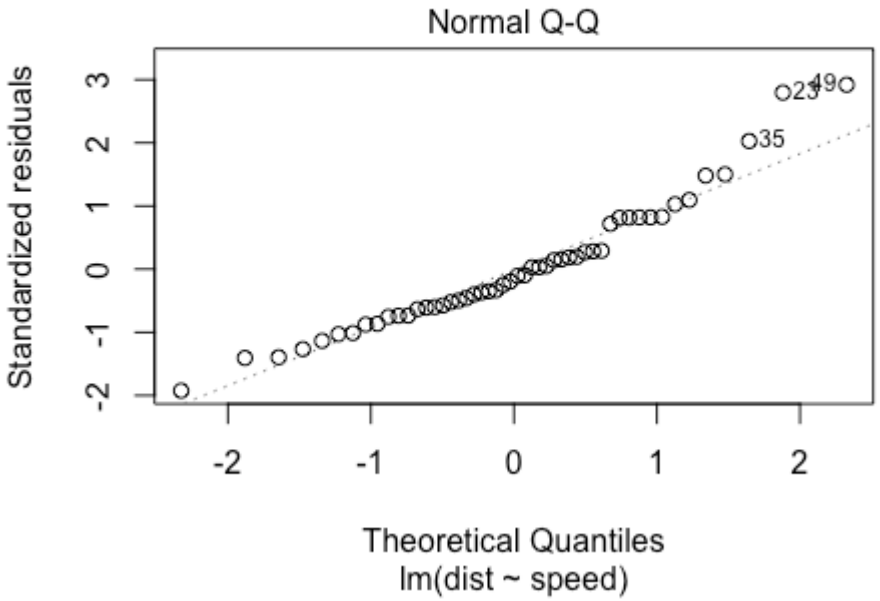
A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.

Importance of QQ Plot in Linear Regression

Most people use them in a single, simple way: fit a linear regression model, check if the points lie approximately on the line, and if they don't, your residuals aren't Gaussian and thus your errors aren't either. This implies that for small sample sizes, you can't assume your estimator $\hat{\beta}$ is Gaussian either, so the standard confidence intervals and significance tests are invalid. However, it's worth trying to understand how the plot is created in order to characterize observed violations.

Let's fit OLS on an R datasets and then analyze the resulting QQ plots.

```
plot(lm(dist~speed,data=cars))
```

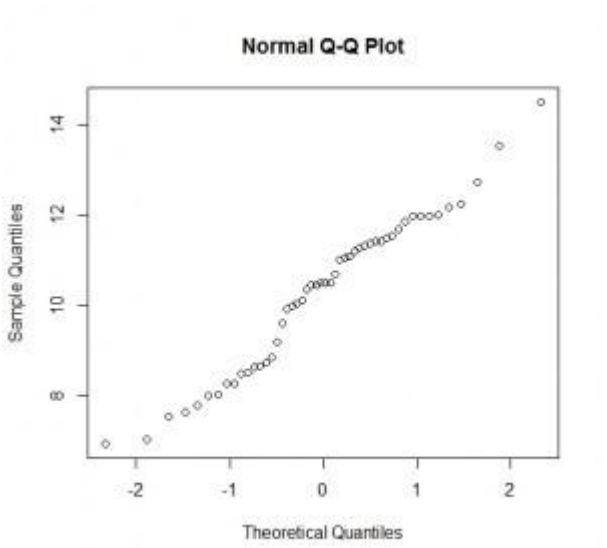


The points approximately fall on the line, but what does this mean? On the x-axis are the theoretical quantiles of a standard normal. That is, we sort the n points, and then for each i, using the standard normal quantile function we find the x so that $P_{\text{std norm}}(X \leq x) = \frac{i-0.5}{n}$. For this dataset, for the case of the leftmost point, we have that $i = 1$ and $n = 50$. Thus

```
qnorm(0.5/50)
[1] -2.326348
```

which looks similar to where the leftmost point is on the x-axis. Intuitively, what this is saying is: we have 50 points and we want their x-values to be such that $P_{\text{std norm}}(X \leq x) = 0.01, 0.03, \cdots, 0.99$. Based on the standard normal distribution, what x do we need to choose? For the y-axis, consider the empirical distribution function of the standardized residuals. We want our corresponding y to be $P_{\text{emp}}(Y \leq y) = 0.01, 0.03, \cdots, 0.99$, but based on the *empirical CDF* of the standardized residuals.

Now how can we characterize the (slight) non-normality? What we see is that on the right hand side of the graph, the points lie slightly above the line. For the very right-most point, this is saying that the value x such that $P(X \leq x) = 0.99$ is larger under the empirical CDF for the standardized residuals than it is under a normal distribution. This suggests a 'fat tail' on the right hand side of the distribution.



Now what are "quantiles"? These are often referred to as "percentiles". These are points in your data below which a certain proportion of your data fall. For example, imagine the classic bell-curve standard Normal distribution with a mean of 0. The 0.5 quantile, or 50th percentile, is 0. Half the data lie below 0. That's the peak of the hump in the curve. The 0.95 quantile, or 95th percentile, is about 1.64. 95 percent of the data lie below 1.64.