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)

First created the dummy variables and calculated the coefficient .Based on that we can infer below.

yr: A coefficient value of '0.2308' indicated that a unit increase in yr variable, increases the bike hire numbers by 0.2308 units.

season_2: A coefficient value of '0.082706' indicated that w.r.t season_1, a unit increase in season_2 variable decreases the bike hire numbers by 0.082706 units.

season_4: A coefficient value of '0.128744' indicated that w.r.t season_1, a unit increase in season_4 variable increases the bike hire numbers by 0.128744 units.

mnth_9: A coefficient value of '0.094743' indicated that w.r.t mnth_1, a unit increase in mnth_9 variable increases the bike hire numbers by 0.094743 units.

weekday_6: A coefficient value of '0.056909' indicated that w.r.t weekday_1, a unit increase in weekday_6 variable increases the bike hire numbers by 0.056909 units.

weathersit_2: A coefficient value of '-0.074807' indicated that, w.r.t Weathersit1, a unit increase in Weathersit2 variable, decreases the bike hire numbers by 0.074807 units.

weathersit_3: A coefficient value of '-0.3070' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable, decreases the bike hire numbers by 0.3070 units.

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

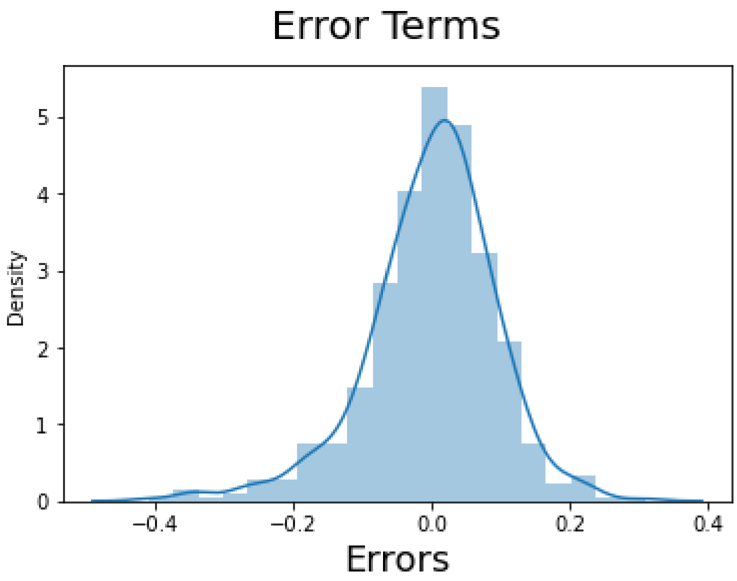
drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

We can see that 'temp','atemp' and 'cnt' has highest correlation with the target variable.

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

We assumed Error terms are normally distributed with mean zero and calculated **y_train_pred** (lr6.predict(X_train_lm6)) i.e for 6th model and Ploted the histogram of the error terms.
From below SS, we could see that the Residuals are normally distributed. Hence our assumption for Linear Regression is valid.



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

As per our final Model, the top 3 predictor variables that influences the bike booking are:

Temperature (temp) - A coefficient value of ‘0.5636’ indicated that a unit increase in temp variable increases the bike hire numbers by 0.5636 units.

Weather Situation 3 (weathersit_3) - A coefficient value of ‘-0.3070’ indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.3070 units.

Year (yr) - A coefficient value of ‘0.2308’ indicated that a unit increase in yr variable increases the bike hire numbers by 0.2308 units.

General Subjective Questions

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

In linear regression, the 2 variables must be linearly co-related.
Mathematically, linear regression equation for n, independent variables and one dependent variable will be: y = b0+b1*x1+b2*x2+…..+bn*xn..
where, b0 is the intercept, that is the condition when dependent variables, does not at all depend on any of the independent variables and b1, b2, ...bn are the coefficients of relationship between the dependent variable and the corresponding independent variable. , i.e. if x1 changes by one unit and all other variables remain unchanged then y changes by b1 units.

Let us consider the equation with one independent variable: y = b0+b1*x.
To find the best fit line, we will use residuals:
residual = yp – y,
where yp is the values of y calculated from the above equation, and y, is the actual value of the dependent data.
So, residual is basically the error in calculation of the actual value vs the predicted value, and therefore our goal is to minimize this error to find the best fit line or the best model.
To achieve this, we will use ordinary least squares method,
For m distinct data points we will have m distinct error points, e1, e2, e3, e4,…em., our goal is to minimize this error using least square method, i.e. e12+e22+e32+...em2=RSS and this, RSS = residual Sum of squares and this should be minimized. After substituting the value of yp in residuals from the equation of line we get RSS = Σ(yi – b0 – b1*x), where I goes from 1 to m.
There are various methods to minimize this error, but generally Gradient Descent is used.

The value of RSS changes with the change in units, for example if x is weight in kg, then the value of RSS will change as the unit of weight changes from kg to grams, therefore, TSS (Total Sum of Squares) is used.
The idea behind TSS is, if I have a linear model where I do not have independent variable, I will only have the intercept, then I will start constructing a model with my intercept's value as the average value of the dependent data point. So, any model that we build on top of that will e better than this model.

R-squared = 1-(RSS/TSS), this explains how good our model is, and how much variance in data we are able to explain from this.

There are a few assumptions while building the linear regression model.
1. Target variable and input variable is linearly dependent.
2. Error terms are normally distributed and are independent of each other.
3. Error terms have constant variance and have mean = 0.

Once the model is built, now we have to check for the significance of the model, i.e. in simple linear regression we have to check for the significance of the value of b1.
So, we will use NULL hypothesis and alternate hypothesis in this case.
H0 => b1=0 and H1 => b1 !=0.

If we can reject the null hypothesis then we can say that the target variable is dependent on the input variables.

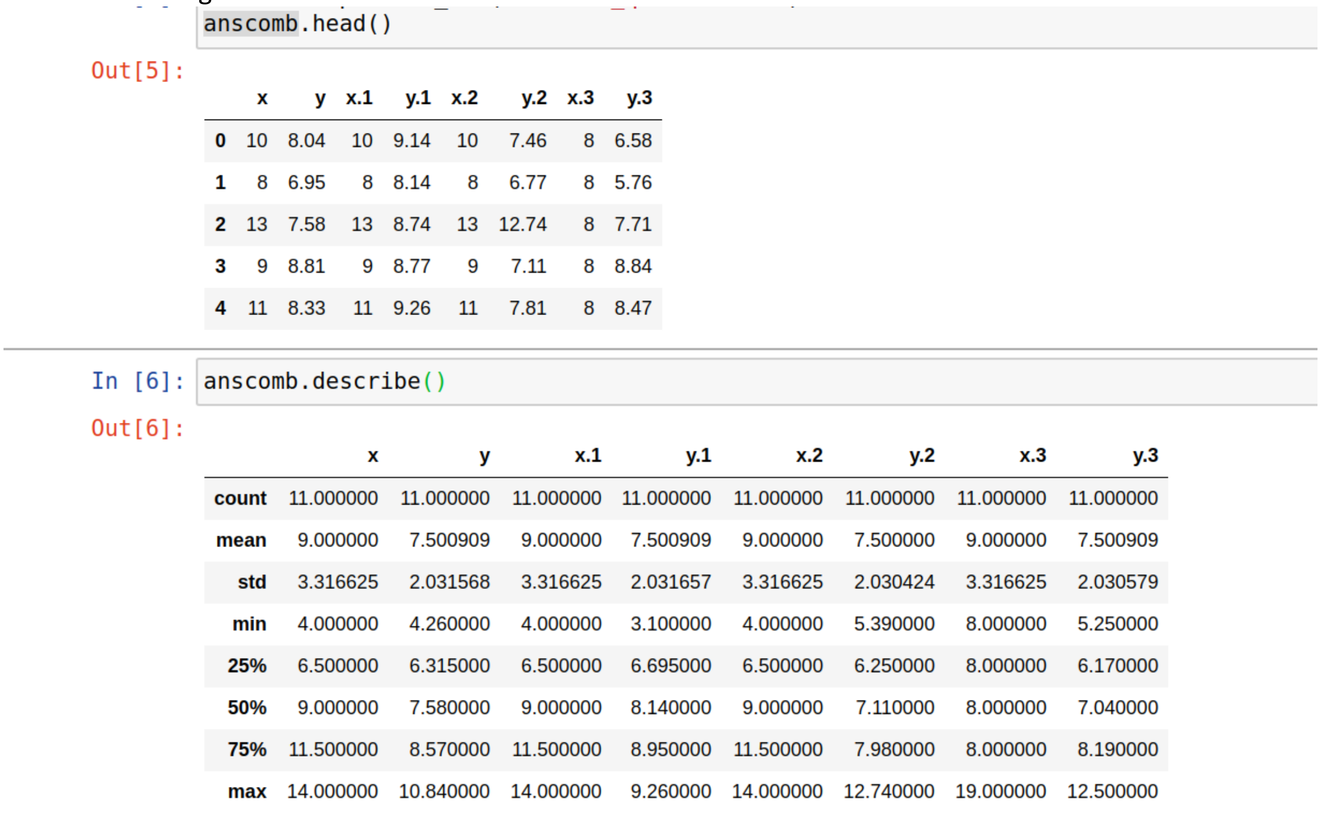
In case of multiple input variables, our equation will become:
$y = b0+b1*x1+b2*x2+.....+bn*xn$..
Apart from all the assumptions and other features that we have in simple linear regression, a new factor of multicollinearity comes into picture, and therefore now we have to check for the Variance Inflation factor, VIF, and based on this value, we will decide which all features will be present in our model, for better fit and which will also avoid overfitting of our model.

$VIFi = 1/(1-Ri)$.

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

- Anscombe’s dataset comprises 4 dataset that have similar descriptive properties, with very different distribution graphs.
- Anscombe’s graph is used to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties

Below image shows the data used for Anscombe and the statistics for the data.



```
anscomb.head()
```

Out[5]:

	x	y	x.1	y.1	x.2	y.2	x.3	y.3
0	10	8.04	10	9.14	10	7.46	8	6.58
1	8	6.95	8	8.14	8	6.77	8	5.76
2	13	7.58	13	8.74	13	12.74	8	7.71
3	9	8.81	9	8.77	9	7.11	8	8.84
4	11	8.33	11	9.26	11	7.81	8	8.47

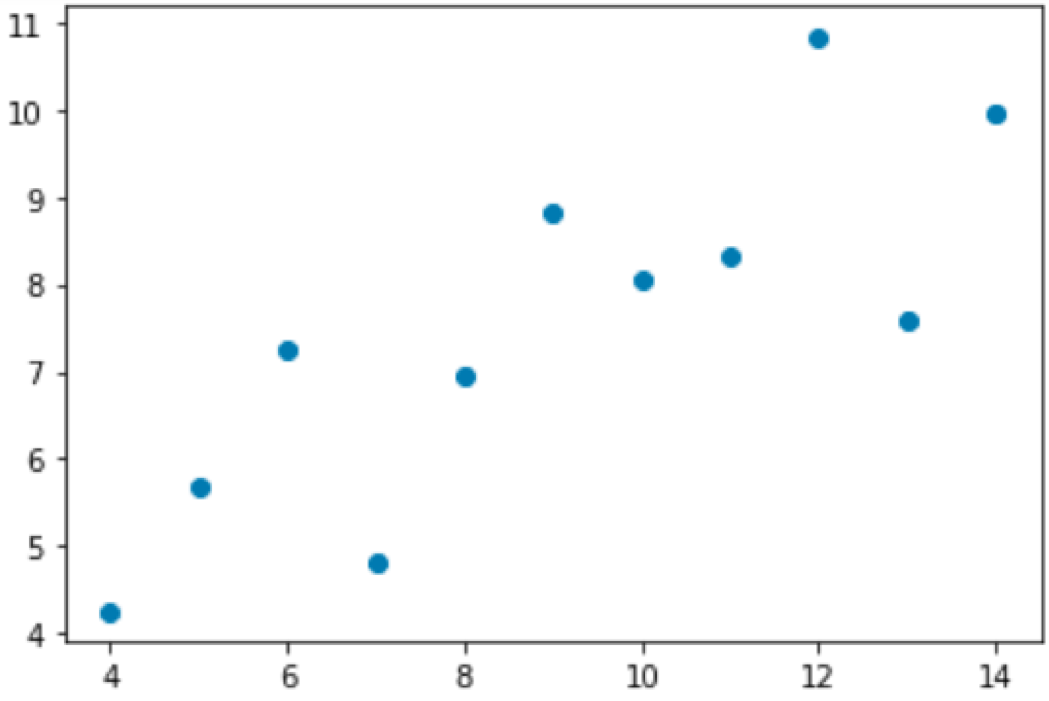
```
In [6]: anscomb.describe()
```

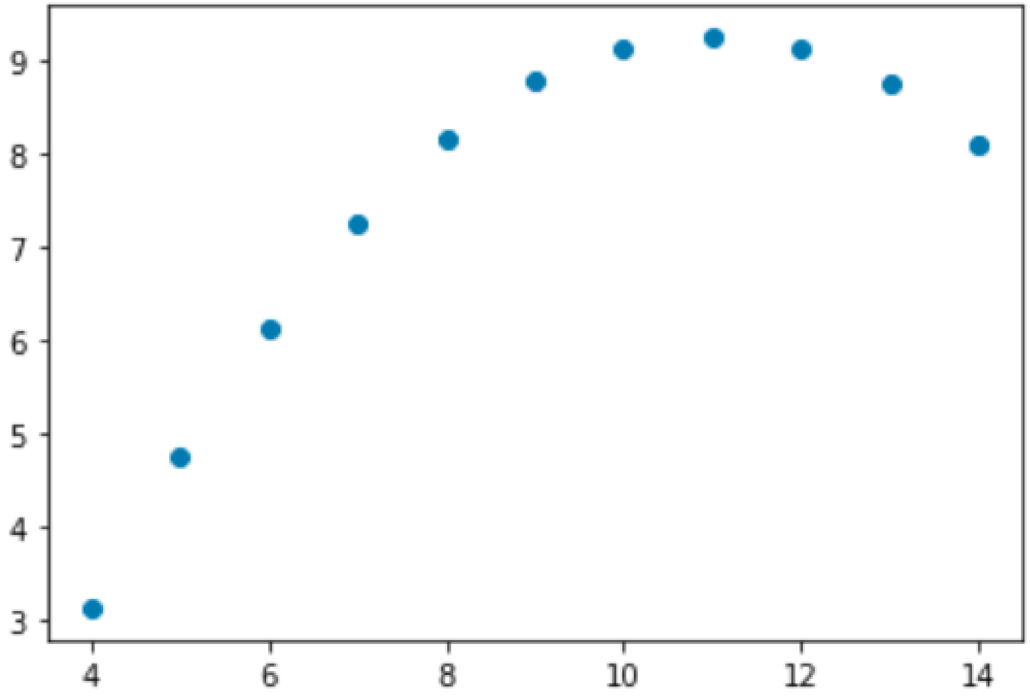
Out[6]:

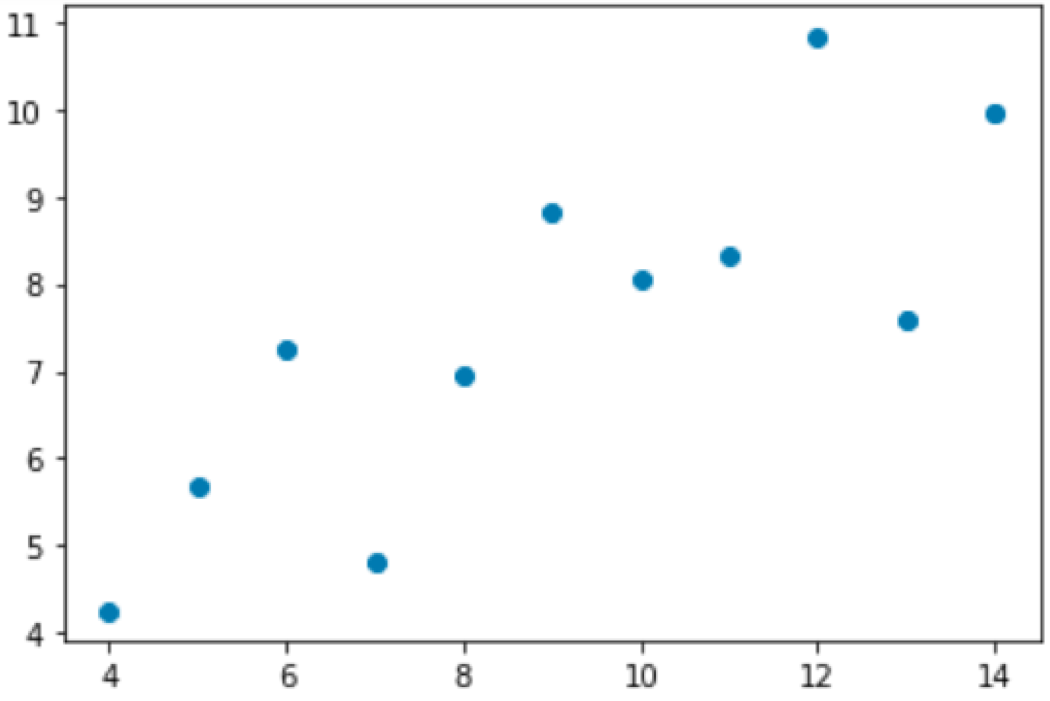
	x	y	x.1	y.1	x.2	y.2	x.3	y.3
count	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000
mean	9.000000	7.500909	9.000000	7.500909	9.000000	7.500000	9.000000	7.500909
std	3.316625	2.031568	3.316625	2.031657	3.316625	2.030424	3.316625	2.030579
min	4.000000	4.260000	4.000000	3.100000	4.000000	5.390000	8.000000	5.250000
25%	6.500000	6.315000	6.500000	6.695000	6.500000	6.250000	8.000000	6.170000
50%	9.000000	7.580000	9.000000	8.140000	9.000000	7.110000	8.000000	7.040000
75%	11.500000	8.570000	11.500000	8.950000	11.500000	7.980000	8.000000	8.190000
max	14.000000	10.840000	14.000000	9.260000	14.000000	12.740000	19.000000	12.500000

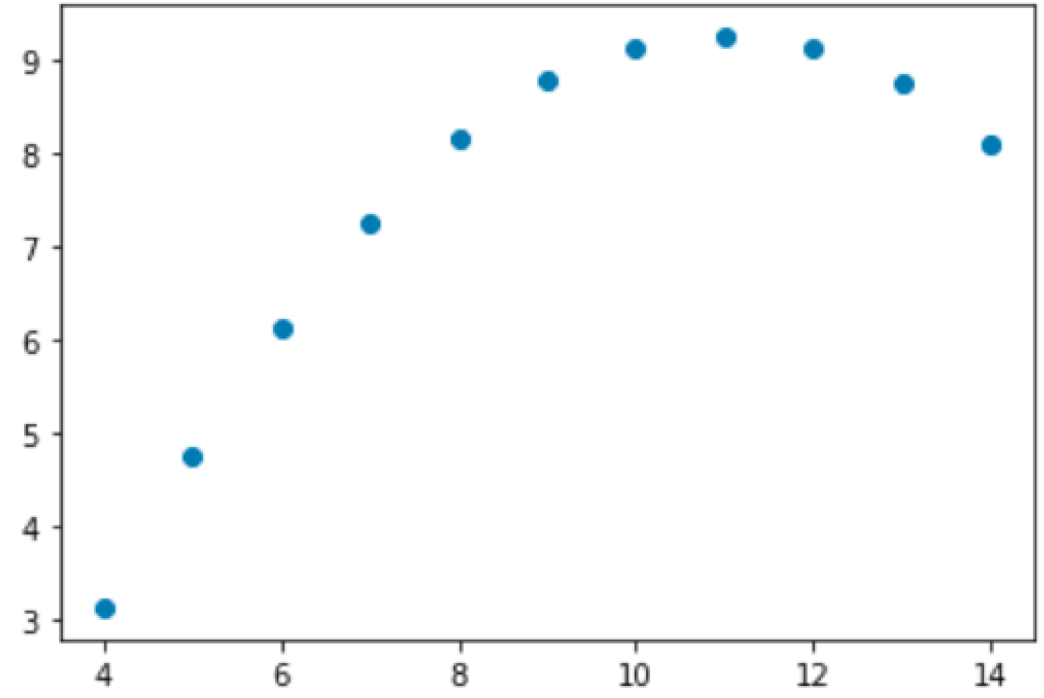
Below 4 graphs represent the relationship between x vs y, x.1 vs y.1, x.2 vs y.2, and x.3 vs y.3.

We see that all 4 x, y pairs have same statistical description but different graphical representation. This is the importance of Anscombe’s quartet.


11
10
9
8
7
6
5
4
4
6
8
10
12
14


9
8
7
6
5
4
3
4
6
8
10
12
14


11
10
9
8
7
6
5
4
4
6
8
10
12
14


9
8
7
6
5
4
3
4
6
8
10
12
14

3. What is Pearson’s R? (3 marks)

Pearson’s R measures the statistical relationship, between two continuous variables and it is based on covariance. It gives information about the magnitude of the association, and the direction of the relationship.
It takes the value between -1 and 1.
In our equation y = b0+b1*x, b1 is Pearson’s R.

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

Scaling is a technique used for standardizing the independent features present in the data in the fixed range. Scaling is performed to bring all the independent variables in same range so that, no variables seem to have major impact on the target variable because of its units.
For example, I have input variables: age, salary, rooms. SO, here we will see that salary variable will have data ranging from 100000 to 10000000. This value is very high in comparison to other two columns age and rooms, so we will use scaling to convert all the input variables in a fixed same range, so that no variables seems to dominate.
Normalized scaling brings all the data in the range of 0 and 1, it also converts outliers in this range.
Standardized scaling brings all the data into a standardized normal distribution with mean =0 and std deviation =1.

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

VIF is infinite => R-squared = 1 => the variable can be expressed exactly as a linear combination of other variables.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.
This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.