Assignment-based Subjective Questions

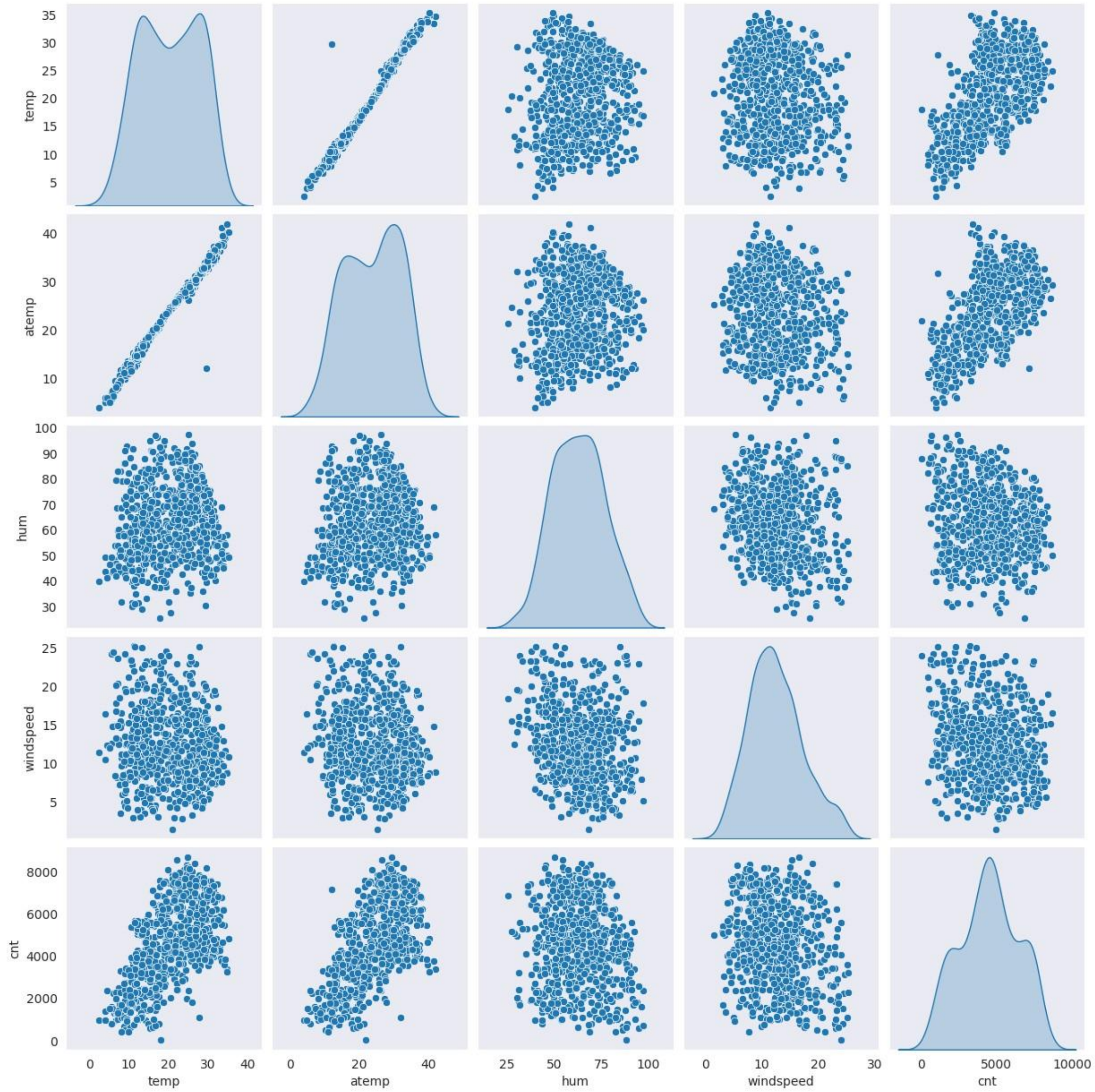
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

- Demand for bikes has **increased** from **2018 to 2019**
- Demand for bikes is **high during the fall**, followed by summer, winter, spring
- Demand for bikes increase with situation of weather. **Better the weather, higher the demand**
- Demand on **working days** is slightly **higher**
- **May to Sept** there is up trend in demand for bikes

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

Using **drop_first=True** helps in reducing the extra column created during the dummy variable creation.Hence it reduces the correlation created among the dummy variables.If our categorical variable has 3 distinct values, using drop_first=True will create only n-1 variable, which is 2 dummy variables.

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

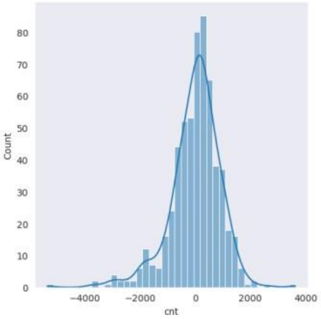


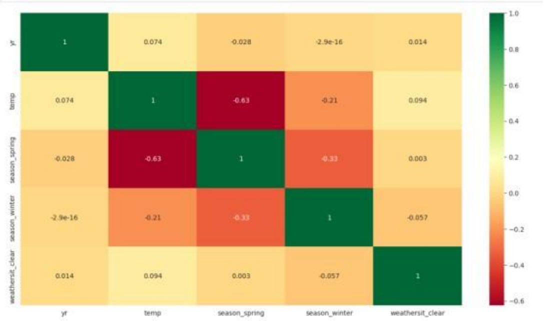
Temp and atemp has equally high correlation with target variable cnt

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

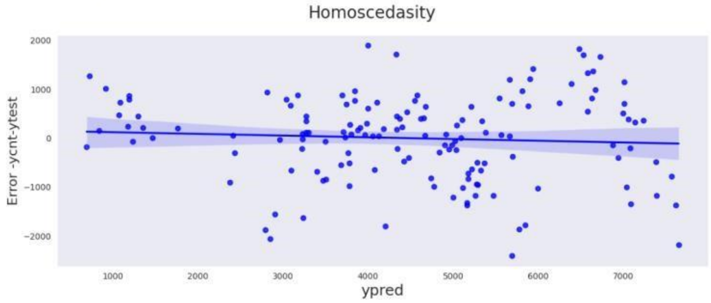
    - Calculated the residuals of the prediction using the model (aka. Error Terms) and created a distplot to check if they are normally distributed with mean centered to zero
    - Checked the correlation analysis of predictor variables used in the model against the target variables and they are under control (0.63) and thus we failed to reject the null hypothesis . We also made sure the VIF factor is < 5 by fine tuning the model predictors

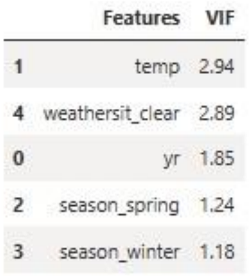
- We checked the homoscedascity of the error terms i.e the constant variance of the error terms using a regression plot and they are inline with our assumption
- Linearity of the predictor variables are checked using a regplot and all the predictors used in our final model follow linearity
-



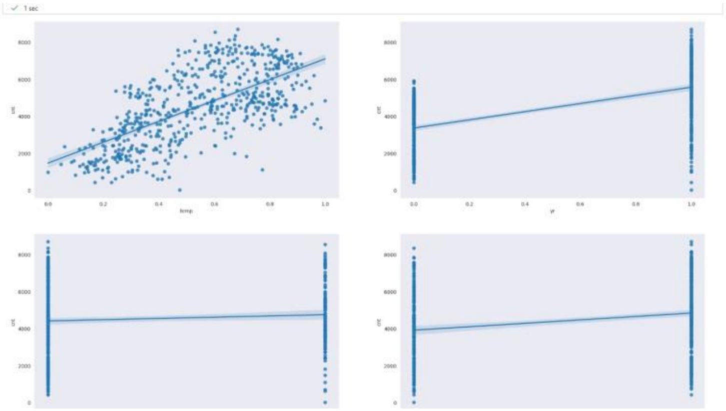


```
Text(0, 0.5, 'Error -ycnt-ytest')
```





	Features	VIF
1	temp	2.94
4	weathersit_clear	2.89
0	yr	1.85
2	season_spring	1.24
3	season_winter	1.18



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

The top 4 influence on the bike booking /demand

- temp : A coefficient value of 4017.586 indicated that a unit increase in temp variable increases the bike demand by same number, provided all other variables are held constant

- season_winter : A coefficient value of 459.0835 indicated that a unit increase in season_winter variable increases the bike demand by same number, provided all other variables are held constant
- weathersit_clear : A coefficient value of 747.949 indicated that a unit increase in weathersit_clear variable increases the bike demand by same number, provided all other variables are held constant
- yr : A coefficient value of 2033.1428 indicated that a unit increase in yr variable increases the bike demand by same number, provided all other variables are held constant

General Subjective Questions

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

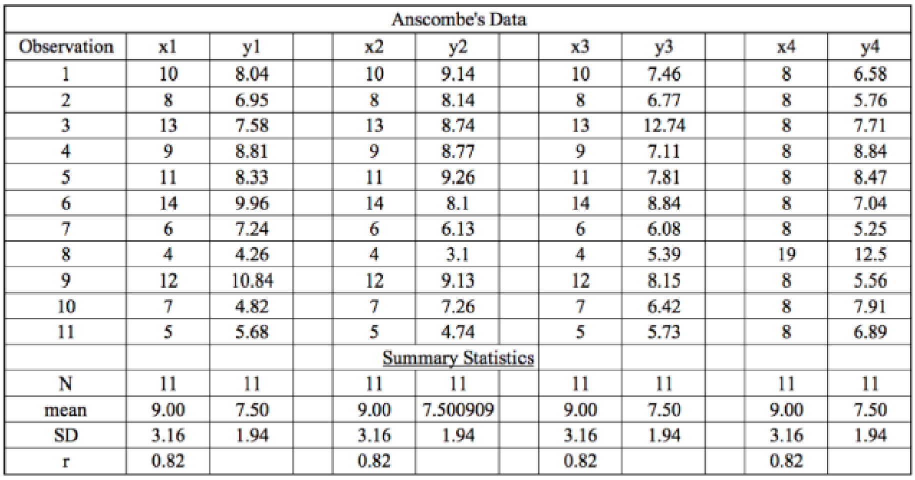
The linear regression uses supervised machine learning technique to statistically establish a relationship between a dependent variable and one or more independent variables. It is used to predict the value of the dependent variable based on the value of one or more independent variables. The relationship of the dependent variable and the independent variable is assumed to be linear in nature.

The goal of linear regression is to find the best fit line that fits the data points. This line is called the regression line. The regression line is defined by an equation of the form y=mx+c where y is the dependent variable and x is the independent variable . m is the slope of the line and b is the y intercept.

There are 2 types of linear regression simple linear regression and multiple linear regression. Simple linear regression involves only one independent variables whereas multiple linear regression involves more than one independent variables. In SLR , we try to fit y=mx+c. In the MLR, we try to fit y = b0 + b1x1+b2x2…+bnxn. Where b0, b1 are the coefficients that represent the contribution of each independent variable. We can use statsmodel.api which uses OLS to best fit line and sklearn and several other libraries to implement linear regression.

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

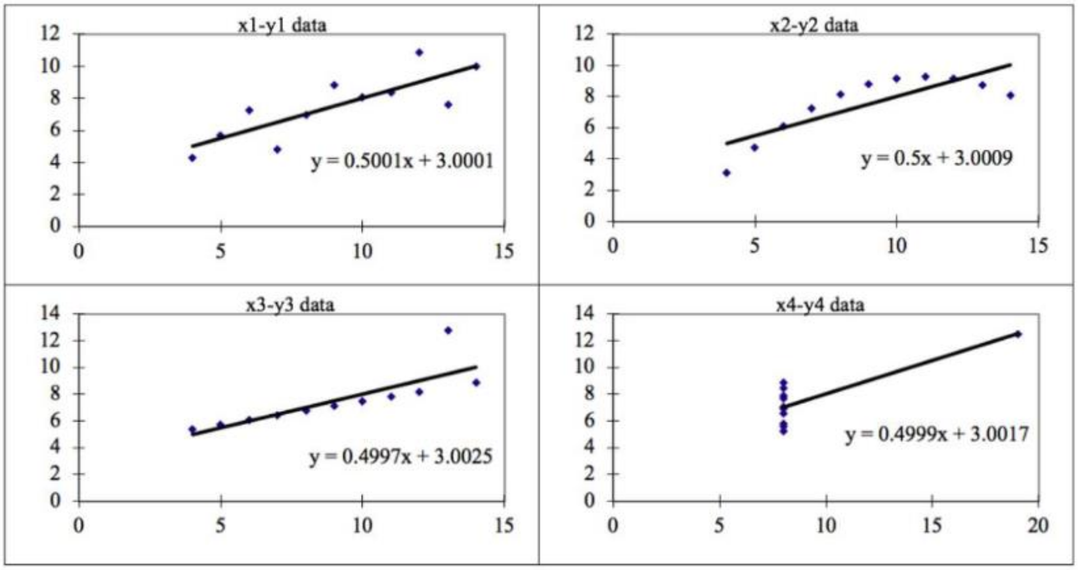
The Anscombe quartet as explained in the data visualization module is a set of four datasets having same descriptive statistics , but when they are plotted on the scatter plot they look different. These datasets are created by Francis Anscombe in 1973 to demonstrate the significance of data visualization. The datasets include 11 pair of x and y values . When we plot these 11 points each plot shows unique variability factors and correlation strength irrespective of they having the same summary statistic (mean, median, mode, quartiles etc). The Anscombe's quartet demonstrates the importance of the EDA, data visualization and shows us why we cannot rely only on the summary statistics to spot the trend, outliers.



Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
					Summary Statistics						
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

Anscombes quartet dataset

Regression line of 4 datasets doesn't fit although they have the same summary statistics



First data set – fits regression line well
Second dataset – doesn't fit the regression line and curvy(non – linear)
Third dataset – contains outliers
Fourth- skewed and outliers

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

Pearsons correlation coefficient (R) is the most common way of measuring the linear correlation. The Pearson correlation varies from -1 to 1. The Pearson correlation measures the strength and the direction of the relationship between two variables. R value from 0 to 1 means positive correlation. R value of 0 means no correlation . R value of 0 to -1 means the correlation is negative.

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

In machine learning, scaling refers to the process of transforming the data so that it fits within a specific scale. Scaling is performed to ensure that all features are on the similar scale.Which can improve the performance of algorithms and the predictions during modelling.
Normalization and standardization are two common ways of scaling. Normalization scales all the numeric variables in the range of 0 -1 , while standardization scales all the numeric variables to have a mean of 0 and standard deviation of 1.
Normalization is useful when we don’t know the distribution of the data or when we know the data does not follow a gaussian distribution.
Standardization is useful when the data is following the gaussian distribution.

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

If 2 predictor variables have perfect correlation then the VIF is infinite. This means one variable can be explained perfectly using a linear relation with other. The value is infinte as the R2 correlation between the variables becomes 1 and the formula for VIF

$1/1-R = 1/0 =$ infinity

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A Q-Q plot (quantile- quantile plot) is a graphical tool for determining if two data sets come from population with a common distribution such as normal, exponential or uniform distribution. It is a probability plot for comparing two probability distributions by plotting their quantiles against each other. In linear regression we use Q-Q plots to check if residuals follow a normal distribution. If the residuals follow a normal distribution, it means that the model is correct and assumptions of linear regression are met. If the residuals do not follow a normal distribution, then it means that there is something wrong with the model and we need to investigate further