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?**

Answer:

a) Demand of bikes is highest in fall season than other three seasons.
b) Demand of bike increased drastically in year 2019 compare to year 2018, is a good sign for business.
c) The trend of bike demand is goes on increasing every month at the beginning of year, it is highest in months May, June, July, August, September and October, then it again falls for the couple of months at the end of year.
d) On holiday, bike bookings are more and bookings are less when there is no holiday.
e) Thursday, Friday, Saturday and Sunday have more demand of bikes than at the beginning of week.
f) Bike demand is almost equal on working and non-working days.
g) People demands bike more on clear_sky weather.

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

Answer:

a) While creating dummy variable, unnecessary column also get created, so in order to drop this column drop_first function is important.
b) Some Correlations are also created among dummy variables, so in order to reduce these correlations we use this function.
c) By default drop_first parameter has Boolean value false, so we have to mention it as True.

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

Answer:

- A ‘temp’ variable has the highest correlation with the target variable ‘cnt’, as there is strong linear pattern is visible
- Also, the coefficient value for temp variable is 0.4777 which is highest among all other predictor variables.

4) **How did you validate the assumptions of Linear Regression after building the model on the training set?**

Answer:

i. **Linear relationship between predictor variables and target variable**
   - By using a heatmap we validate this linear relationship

ii. **Normal distribution of error terms**
   - First we have to find error values and plot these error values on distribution plot to check normality of error terms with **mean zero**

iii. **Error terms are independent of each other**
   - We identify the pattern between error values and predicted values of target variable(by plotting regplot)
   - There is no any specific pattern seen so this validates independency of error terms.

iv. **Homoscedasticity**
   - Plotting a scatter plot between residuals and count, we observe that no low or high concentration of data at any specific region.
   - Error terms should have constant variance values

v. **Multi-colinearity**
   - I observe that, VIF values for predictor variables of final model are less than 5
   - After drawing heatmap of predictor variables of final model, it seems that there is no multicolinearity between variables

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

Answer:

i. **Temp**
   - Since the coefficient value for temp is 0.4777 which is highest among all other predictor variables.
   - There is strong linear relationship between temp and a target variable(cnt), as compare to other predictor variables

ii. **Winter (season)**
   - Since the coefficient value for winter (season) is 0.0945 which is second highest among all other predictor variables.

iii. **Sept(month)**

- Since the coefficient value for sept (month) is 0.0910 which is third highest among all other predictor variables.

Above three are the top features contributing significantly towards explaining the demand of bikes respectively

General Subjective Questions

1) Explain the linear regression algorithm in detail.

Answer:

- Linear regression is a statistic model to analyse the linear relationship between target variable(dependent variable) and predictor variables (independent variables)
- Linear regression has two types, simple linear regression and multiple linear regressions.
- Simple linear regression represented by equation: $y = \beta_0 + \beta_1 X$
  Where: y is target variable (dependent)
  X is predictor variable (independent)
  β_1 is slope of linear regression line
  β_0 is constant or y-intercept
- Multiple linear regression represented by equation:
  $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 ....+ \beta n Xn$
  Where: y is target variable (dependent)
  X_1, X_2, X_3 are predictor variables (independent)
  β_1, β_2, β_3 are coefficients of X_1, X_2, X_3 respectively
  β_0 is constant or y-intercept
- In multiple linear regression model fits **hyperplane** instead of line in simple linear regression
- In multiple linear regressions, coefficients are obtained by minimizing sum of squared error.
- For inference, assumptions for linrear regression are: normal distribution of error terms (zero mean), independency of error terms, constant variance of error terms (homoscedasticity), linearity between dependent and independent variables.

2) Explain the Anscombe's quartet in detail.

Answer:

- Anscombe's quartet was developed by statistician Francis Anscombe
- It consist of four datasets, each contains eleven(x,y) pairs
- Every dataset shares same descriptive statistics, but when they are graphed things changes totally
- Following image shows Anscombe's quartet:

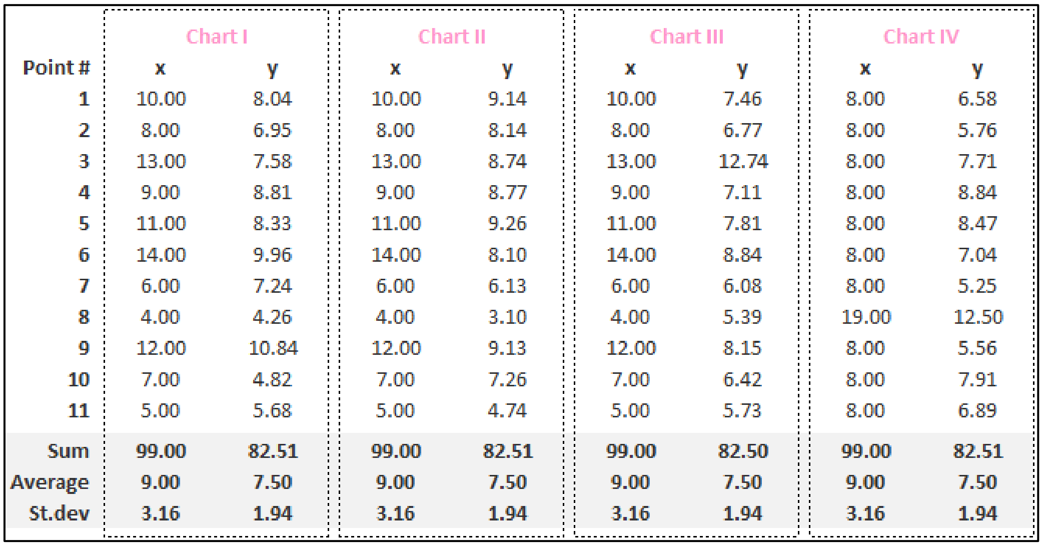
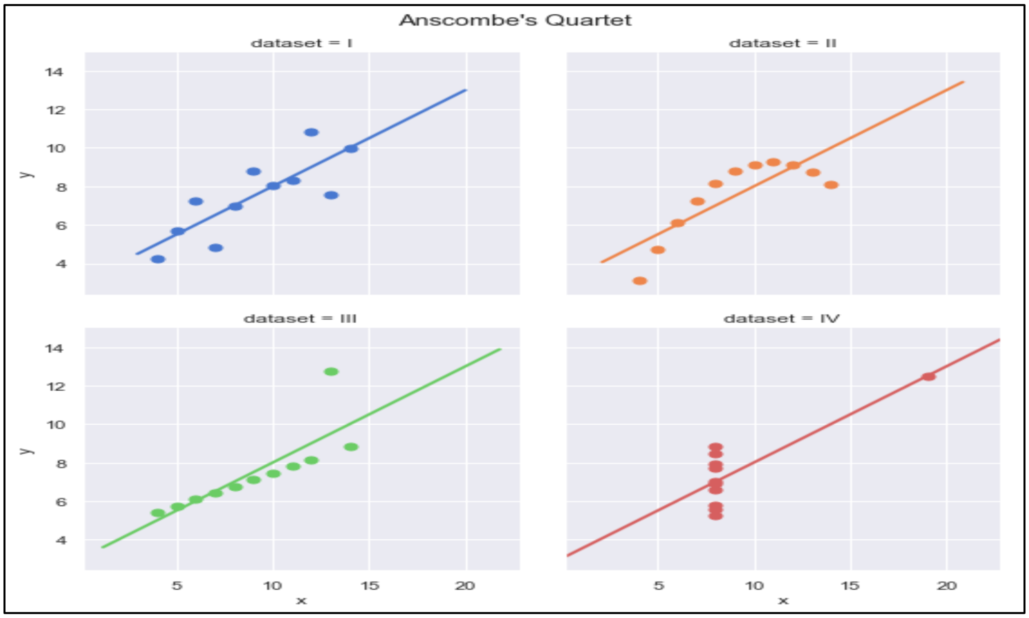


	Chart I		Chart II		Chart III		Chart IV	
Point #	x	y	x	y	x	y	x	y
1	10.00	8.04	10.00	9.14	10.00	7.46	8.00	6.58
2	8.00	6.95	8.00	8.14	8.00	6.77	8.00	5.76
3	13.00	7.58	13.00	8.74	13.00	12.74	8.00	7.71
4	9.00	8.81	9.00	8.77	9.00	7.11	8.00	8.84
5	11.00	8.33	11.00	9.26	11.00	7.81	8.00	8.47
6	14.00	9.96	14.00	8.10	14.00	8.84	8.00	7.04
7	6.00	7.24	6.00	6.13	6.00	6.08	8.00	5.25
8	4.00	4.26	4.00	3.10	4.00	5.39	19.00	12.50
9	12.00	10.84	12.00	9.13	12.00	8.15	8.00	5.56
10	7.00	4.82	7.00	7.26	7.00	6.42	8.00	7.91
11	5.00	5.68	5.00	4.74	5.00	5.73	8.00	6.89
Sum	**99.00**	**82.51**	**99.00**	**82.51**	**99.00**	**82.50**	**99.00**	**82.51**
Average	**9.00**	**7.50**	**9.00**	**7.50**	**9.00**	**7.50**	**9.00**	**7.50**
St.dev	**3.16**	**1.94**	**3.16**	**1.94**	**3.16**	**1.94**	**3.16**	**1.94**

- From above fig, it seems that all 4 datasets have same mean and std. deviation.
- But, when we plot these 4 datasets on x-y plane, it shows same regression line but different data points are there, as shown below:

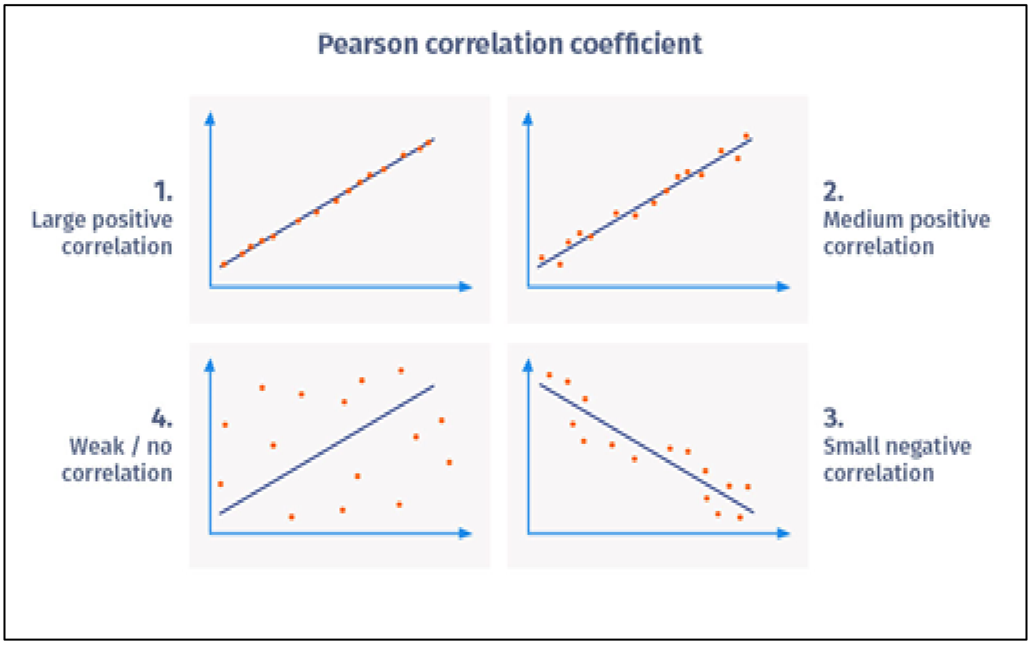


- Dataset – I shows well fitted linear model
- Dataset –II shows there is no normal distribution of data points
- Dataset – III shows, one outlier
- Dataset – IV shows one outlier is enough to produce high correlation coefficient.
- Anscombe's quartet is very important for visualization of data in data analysis

3) What is Pearson's R?

Answer:

- Pearson's R measures strength of linear relationship between two variables
- Pearson's R value always lie between **-1 to +1**
- When Pearson's R value is zero, it means there is no any correlation between two variables.
- When Pearson's R value is above zero (up to +1), it means there is positive correlation between two variables.
- When Pearson's R value is below zero (up to -1), it means there is negative correlation between two variables.
- When Pearson's R value is maximum (+1), it means there is large positive correlation between two variables.
- When Pearson's R value is minimum (-1), it means there is small nagative correlation between two variables.



4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:

- Scaling is a technique to standardize independent features present in the data within fixed range
- Why do we need to scale features:
  - Ease of interpretation
  - Faster convergence for gradient descent methods(gradient descent is an algorithm used to minimize cost function)
  - After scaling only coefficient (feature) changes
- There are two scaling methods:
  - MinMax scaling(normalizing)
  - standardization

MinMax scaling (Normalising)	Standardization
It is used when features are of different scales	It is used when we want to ensure zero mean and unit std. deviation
Minimum and maximum values of features are used for scaling	Mean and std. deviation are used for scaling
It scales between values [0,1] or [-1,1]	There is no certain range of values
It takes care of outliers	It is not much affected by outliers
For normalization, Scikit package provides a transformer MinMaxscaler	For standardization, Scikit package provides a transformer StandardScaler

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

Answer:

- VIF is measure of, how well a predictor variable is correlated with all other variables excluding a target variable.
- Hence, when VIF is infinite means, there is very strong (perfect) correlation between two independent variables.

- Perfect correlation means R-squared value is equal to 1, which is the condition of over fitting (multicollinearity)
- Hence, VIF value if infinite because of multicollinearity or over fitting
- To solve this problem of multicollinearity we have to drop a variable from data set which causes this.

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

Answer:

- Q-Q plot is a graphical technique to determine two datasets from same population have common distribution or not.
- **Use of Q-Q plot:**
  - ➢ Q-Q plot is plot of quantiles of first data set verses quantiles of second data set
  - ➢ Quantile represents what percentage of data points fall below of it and above of it, means for example, 0.4 quantile is a point at which 40% of data falls below of it and remaining 60% data falls above it.
  - ➢ Then the 45° reference line also plotted
  - ➢ When the two datasets from same population having common distribution then these all points should fall approximately along this reference line.
  - ➢ But if all the points of two data sets does not fall approximately along this 45° reference line implies that these two datasets coming from same population having different distributions(i. e no common distribution)
- **Importance of Q-Q plot:**
  - ➢ To justify the assumption that, two data sets from same population having common distribution, Q-Q plot is important
  - ➢ If there is common distribution, then location and scale estimators can pool both data sets to obtain common scale and location
  - ➢ If there is no common distribution between to data sets, then also Q-Q plot is useful to understand the difference