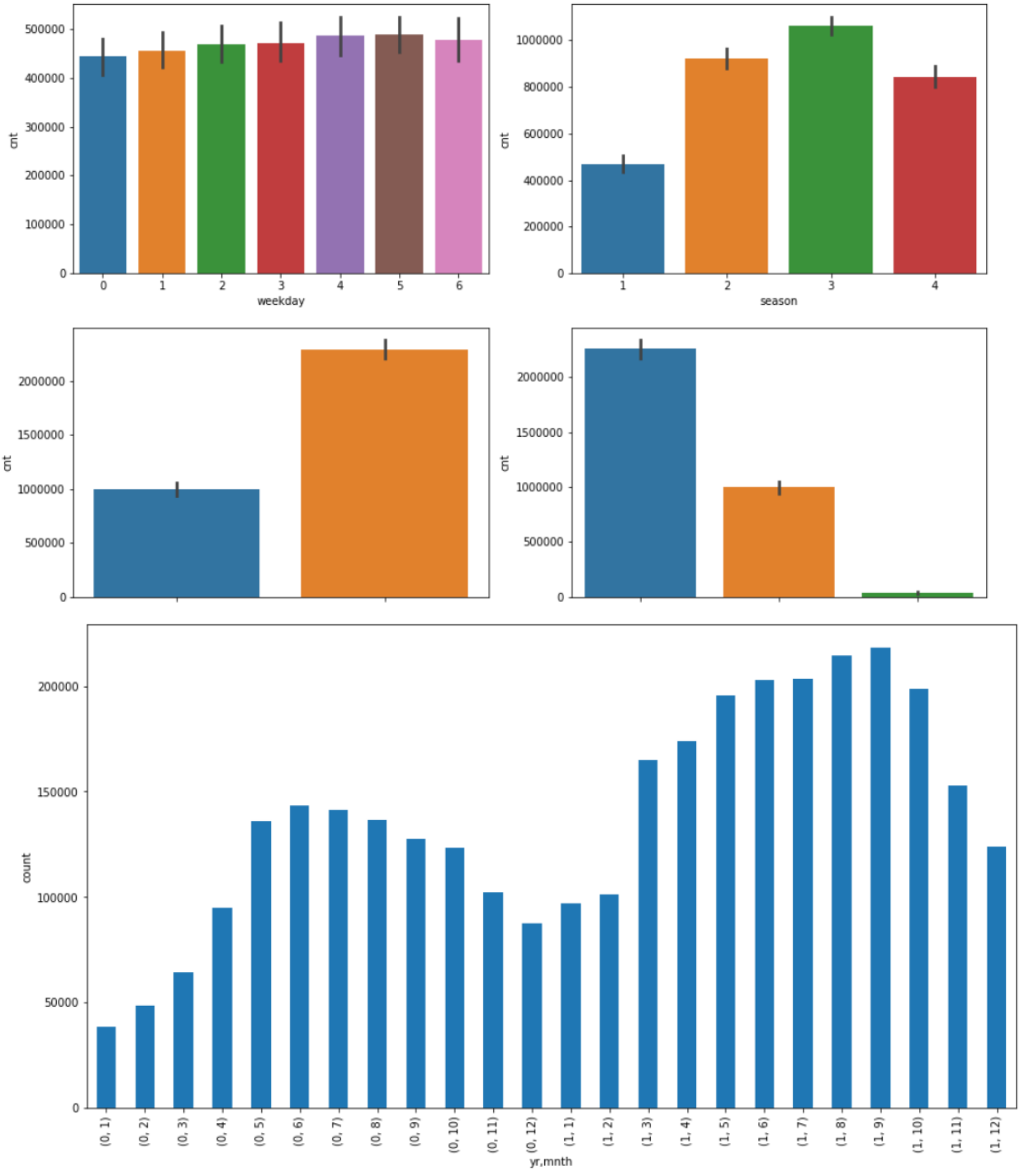
Assignment Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

A1. Key Observations from categorical variables are as follows

- During Fall (season), highest number of bookings are recorded.
- Bookings are very high on working days then on non-working days. People prefer to use bike service for travelling to work.
- People are more likely to use bike service on a clear or partly cloudy day. and hardly use the service during snow season.
- The last figure below shows that over the two years the demand for bikes has increased.
- This also confirms that during winters i.e. months January, February, November, December demand decreases significantly.



Q2. Why is it important to use drop_first=True during dummy variable creation?

A2. If we don't use drop_first=True we will get a redundant variable. For example, in our data set. Holiday column is a categorical column.

Row	**Holiday**
0	0
1	0
2	1

After applying get dummies without drop_first
pd.get_dummies(df.Holiday, prefix='Holiday')

Row	Holiday_0	Holiday_1
0	1	0
1	1	0
2	0	1

Both the variables are clearly strongly correlated leading to multi-collinearity.
In general, among k categories – kth category can always be shown by setting k-1 categories as 0 or absent.
In context of Linear Regression Model, multi-collinear parameters are not advisable as it affects the interpretation and inference derived from the model.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

A3. Temperature (temp or atemp) is strongly correlated with target variable **cnt.**
(Both atemp and temp are highly correlated among themselves so both can be considered)
Structure of plot for temp and atemp is also similar w.r.t. target variable count.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

A4. Assumptions of linear regression are as follows:

- Linear relationship between X and Y variables.
- No multicollinearity among predicting variables
- Error terms are normally distributed.
- Error terms are independent of each other.
- Error terms have constant variance. (homoscedasticity)

Verification of assumptions on data set.

1. **Linear relationship between X and Y**.- We have plotted pair plots for numerical variables. It is quite evident that numerical variables have linear relation with target variable with different levels of variances.
2. **No multicollinearity among predicting variables**- This can be detected in 2 ways:
    a. **Scatter Plots** – We can visually detect any relation between the variables.
    b. **Variance Inflation Factor (VIF) -** VIF = $1/(1-R^2)$. It is calculated by building model among predicting variables and determine if any variable can be predicted by other variables.

i. **VIF<5 -** Variable is sufficiently independent
ii. **VIF>5:** Variable may or may not be independent. Worth further analysis
iii. **VIF>10:** Variable is not independent of others and should be either treated or eliminated.

3. **Error Terms are normally distributed –** We perform residual analysis on predicted values in training dataset. Plotting a dist plot can help in visually verifying this.
4. **Error terms are independent of each other –** This can be again done using following ways
    a. **Scatter plots -** There should be no visible pattern in the error terms.
    b. **Durbin Watson Test-** Test statistic obtained in this test has value from 0-4
        i. If value = 2 - no autocorrelation
        ii. 0 to <2 - positive correlation
        iii. >2 – 4 - negative correlation
5. **Error terms have constant variance. (homoscedasticity) –** This can be again identified from scatter plot of error terms. The variance in terms should not change across different subsamples. The Goldfeld-Quandt Test can also be used to test for heteroscedasticity.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

A5. Top 3 features contributing to explain the demands of shared bikes are:

1. weathersit_3
2. year(yr)
3. season_spring (season_1)

P.T.O.

General Subjective Questions

Q1. Explain the Linear Regression algorithm in detail.

A1. Linear Regression is a machine learning algorithm based on supervised learning. It is used for regression tasks i.e. predict a target numerical (or continuous) variable based on some independent variables. It basically involves mapping an output variable Y to a function F (x_1, x_2, x_3, …, x_N) where F is a linear function of n variables.

Input parameters- A dataset of n statistical units.

Output - Result of the Linear Regression Algorithm is best fit line/hyperplane (for MLR).

Assumptions of linear regression are as follows:

- Linear relationship between X and Y variables.
- No multicollinearity among predicting variables (MLR)
- Error terms are normally distributed.
- Error terms are independent of each other.
- Error terms have constant variance. (homoscedasticity)

Hypothesis function for linear regression – $F(y_i) = \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \theta_3 x_{3i} + \cdots + \theta_n x_{ni}$ - eq.1

θ_0 – intercept θ_1 – coefficient of x_1 … θ_n – coefficient of x_n

Linear regression tries to find the values of θ_0 θ_1 θ_2 θ_3 θ_n by minimizing the cost function.

Cost function for Linear Regression - $J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$.

This function is also known as Mean Squared Error (MSE).

Using the MSE we start with certain values of these variables and keep adjusting their values till MSE settles at the minima. This technique of minimization is called Gradient Descent.

The values of the coefficients (θ_0 θ_1 θ_2 θ_3 θ_n) obtained via this minimization depict the best fit line/plane for the given data. This can be substituted in eq.1 to make predictions for corresponding y values.

We perform above steps on the training data set. After this we perform hypothesis testing to understand the significance of different variables.

H_0: Variable is not significant.

H_1: Variable is significant.

Variables with high p-values are removed and cost function is reassessed.

We also calculate variance inflation factor (VIF) for each variable after every step. Variables with high VIF are also removed. This helps in tackling multicollinearity issue.

Once we have achieved the required minimization, we perform the prediction testing on the test data and validate our assumptions.

Q2. Explain the Anscombe's quartet in detail.

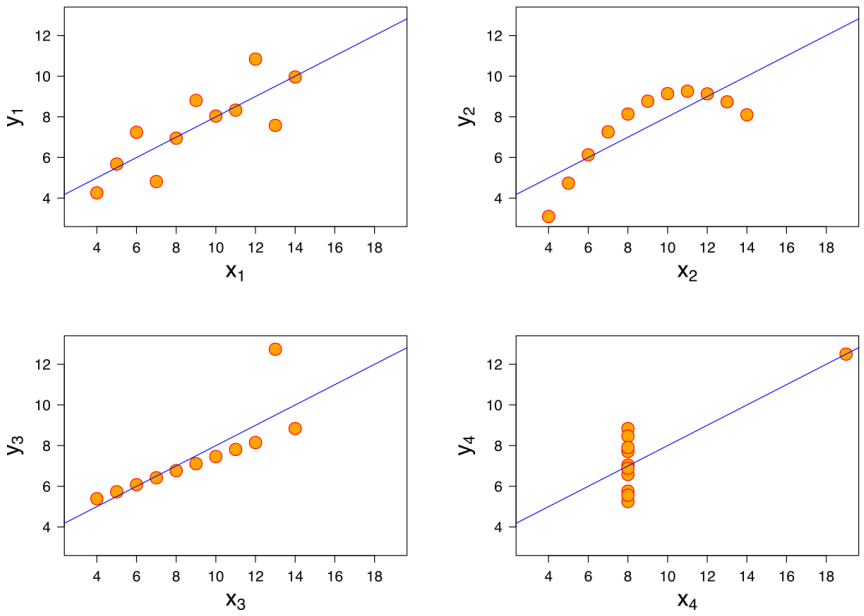
A2. Anscombe's quartet was developed by statistician Francis Anscombe. It helped explain the concept that only mathematical value like mean, variance, standard deviation is not enough to get complete picture of the data. Visual analysis plays important role in understanding of the data.

It consists of 4 different data sets with exactly same statistical values. However, when these data sets are graphed, they show a totally different picture.

Source: Dataset

	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

Source: Anscombe Data Set Plot



Observations:

1. Dataset 1 is clean and well fitting for linear models
2. Dataset 2 cannot be analyzed using linear regression due to curvature of the plot. It does not justify the first assumption of linear regression.
3. Dataset 3 – It has linear distribution largely. But it is the outlier which resulted in mathematical values matching the above 2 graphs
4. Dataset 4 – this is like dataset 3. If outlier is not present the values for statistical properties will be entirely different. Presence of outlier is enough for producing high correlation coefficient.

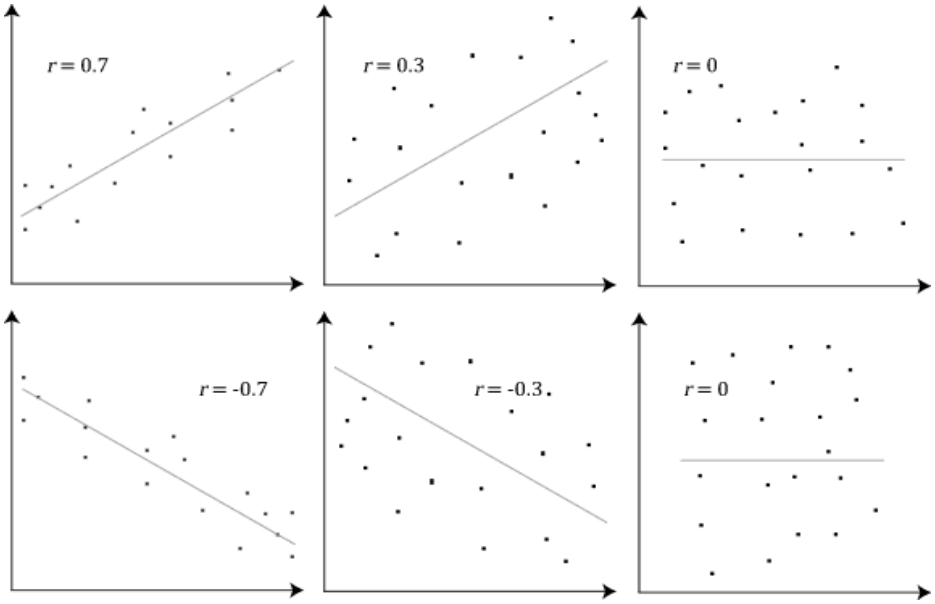
Q3. What is Pearson's R?

A3. Pearson's R or Pearson's product momentum correlation coefficient is the measure of strength of linear association between two continuous variables.

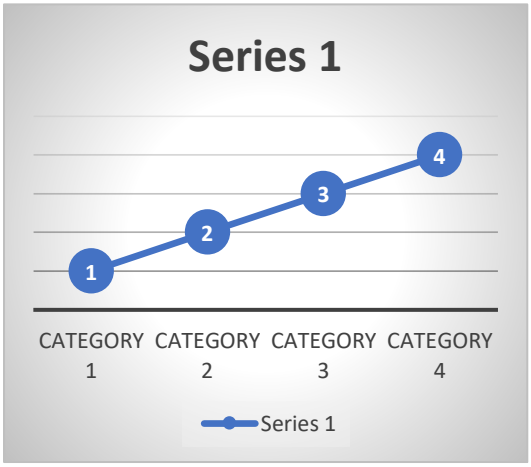
Pearson's R helps in understanding the variance of datapoints w.r.t the best fit line i.e. how far are the data points from the best fit line that is drawn to explain the relation between two variables.

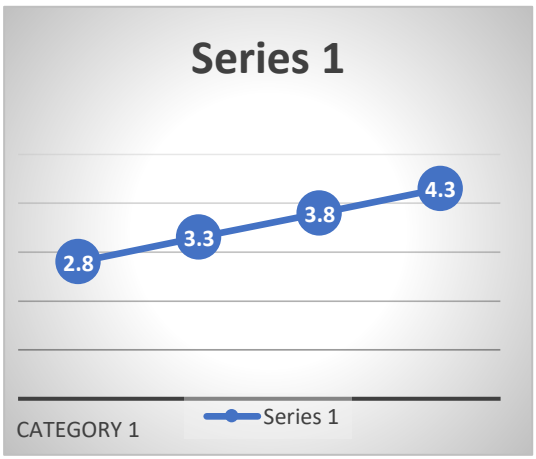
Pearson's R can take values from -**1 to 1.** Value of -1 or 1 means that all the data points are along the best fit line. This means that there is no variation in the data points away from this line.

If value is 0 it indicates maximum variation around the best fit line. Image Source



It is worth noting that value of Pearson's R does not tell any thing about slope of the line. It just describes how the points are distributed around the line.





Both the lines have different slopes but same value for coefficient i.e. 1

- Value = 0 indicates no correlation among the variables
- Value > 0 indicates positive correlation i.e. if one variable increases other also increases
- Value < 0 indicates negative correlation i.e. if one variable decreases other increases and vice versa.

In general values in ranges [-1, -0.5] and [0.5, 1] indicate strong correlation.

Values in range (-0.5, -0.3] and [0.3, 0.5) indicate medium correlation.

Values in range (-0.30) and (0, 0.3) indicate small correlation. 0 as mentioned above indicates there is no correlation

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

A4. Feature Scaling is a technique used to standardize the features present in a data in a fixed range. If the variables in the data set are in different range the variable with larger numerical values (irrespective of units) will dominate the result of the algorithm.

For example – Consider age and income variables in a data set. We observe the following.

Age will generally range from say 30-60, income can range from 2,00,000 – 10,00,000. Now both these variables are not correlated.

Consider a classification problem with 2 categories. We need to predict the category of a new data point.

Assume that algorithm using Manhattan Distance method i.e. it calculates the sum of absolute difference between existing and new points to make the prediction. P1 = (20, 500000) P2 = (57, 800000)

Distance = (|40-57|+|5,00,000-8,00,000|).

Clearly the income variable will dominate the result.

In linear regression problems, gradient descent algorithm is impacted by the scale. If we rescale the features in common range the gradient descent algorithm works faster.

Scaling the variables will bring both age and income variables in comparable ranges. Due to this the result will not be impacted by the range or magnitude of one variable but by the contribution it makes in prediction.

Hence scaling is required.

Normalized scaling uses the following equation to re-scale the variable with distribution **between 0 and 1.**

$$X_{new} = \frac{X_i - Min(X)}{Max(X) - Min(X)}$$

Standardized Scaling uses following equation to re-scale a variable so that the result has a **mean value = 0 and variance = 1**

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

Since normalization brings every value in range of 0 and 1, outliers will be impacted. Standardized scaling does not have a bounding range, so the outliers are not impacted by standardization.

Q5 You might have observed that sometimes the value of VIF is infinite. Why does this happen?

A5. VIF = $1/(1-R^2)$. Mathematically for VIF = infinite.

$1-R^2 = 0$

$R^2 = 1$ => R= +1 or -1

This would be the case if the variable under question is directly correlated to some other variable. E.g. temp and atemp in our assignment are very highly correlated.

Date of birth and Age are 2 highly correlated values. One can be directly derived from other. In this case VIF value will tend to infinity.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A6. Q-Q plot stands for Quantile-Quantile Plot. It is graphical method to determine if two data sets come from population with similar distribution. For example, we can find a given sequence of numbers is normally distributed by plotting a Q-Q plot with standard normal distribution.

It is a plot of quantiles of first dataset against the quantiles of the second data set.

It can be used to detect following characteristics about the data sets.

- If the 2 datasets come from **same population**, the points on this plot should fall on an approximately straight line with an angle of 45 degree from x-axis.
- Have common location and scale
- Have similar distribution shapes
- Have similar tail behavior

X- Axis – Quantiles of 1 dataset

Y- Axis- Quantiles of 1 dataset

In linear regression, Q-Q plot it can be used to visualize if residual errors are normal in nature. For example, here is the plot from our bike assignment.

```
from statsmodels.graphics.gofplots import qqplot
qqplot(lm.resid, line='s')
```