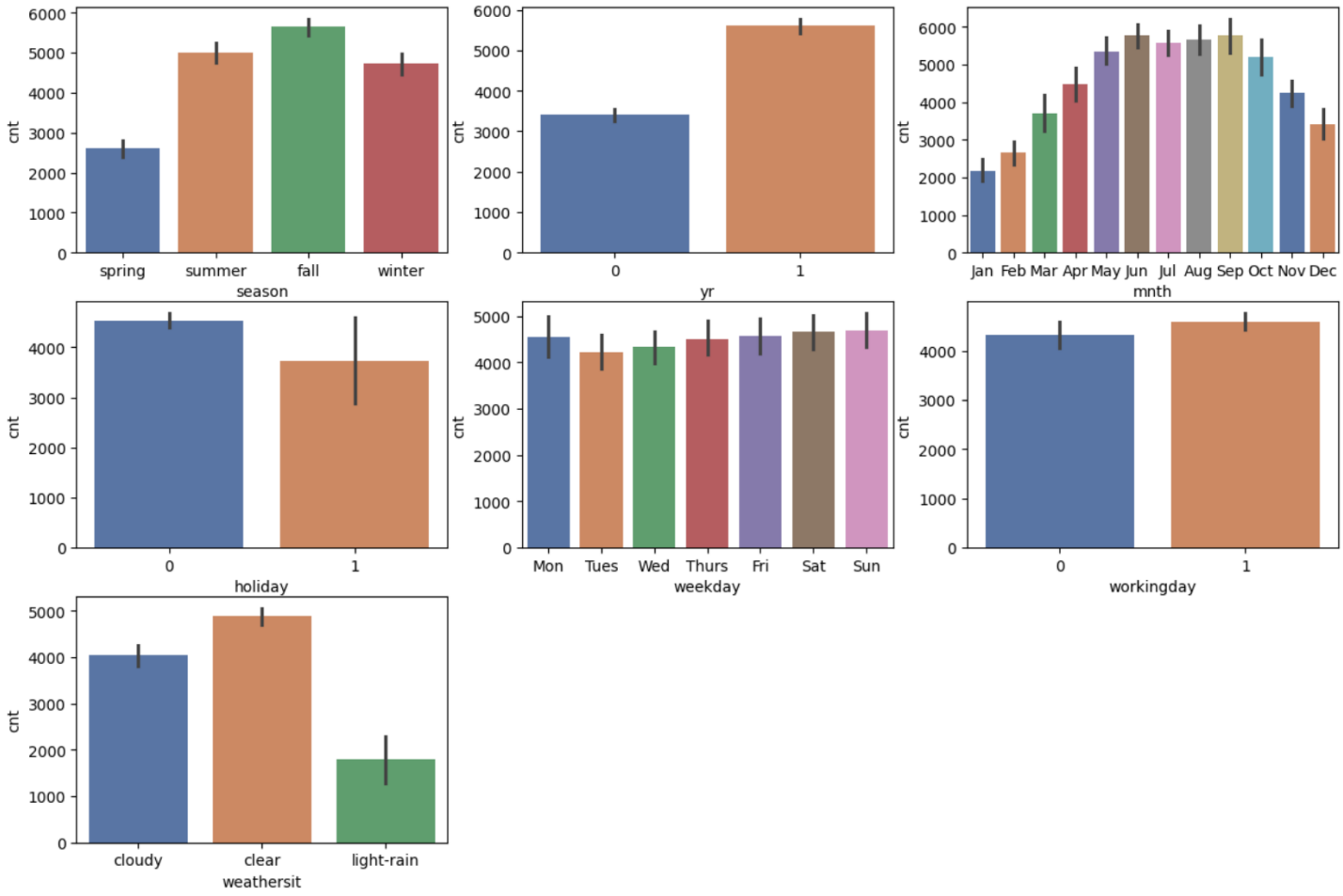
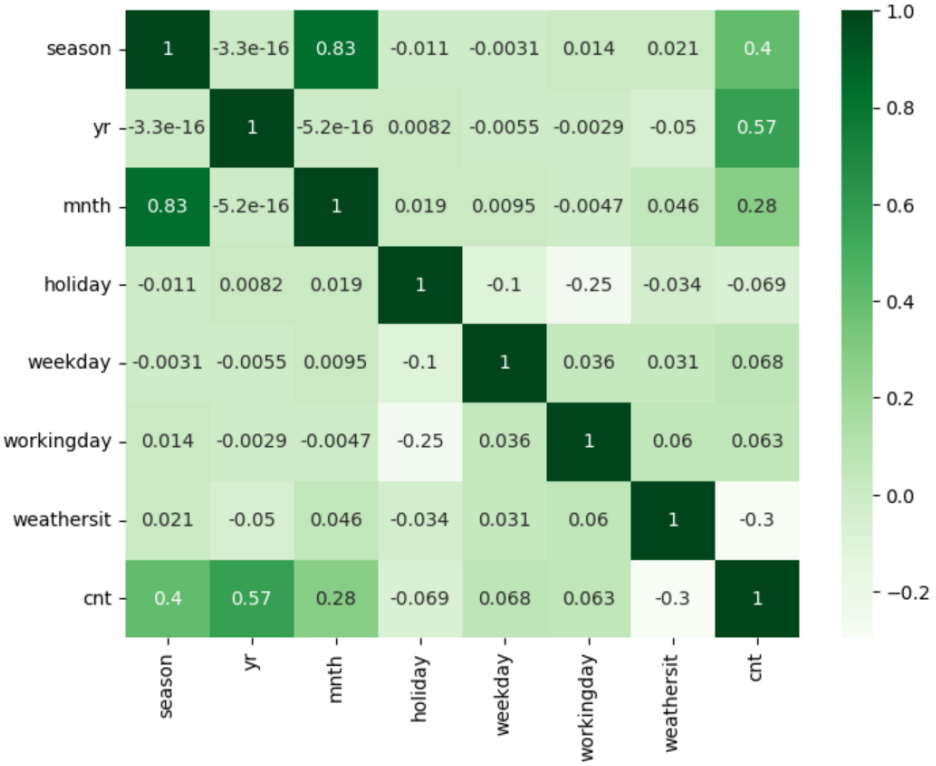
Answers to "Assignment-based Subjective Questions" :

Ans 1 : To analyze the impact of categorical variables on the target variable, let us look at the following plot with various categorical variables on X-axis and 'cnt' on Y-axis.



Let us also look at the heatmap of the categorical variables & target variable :



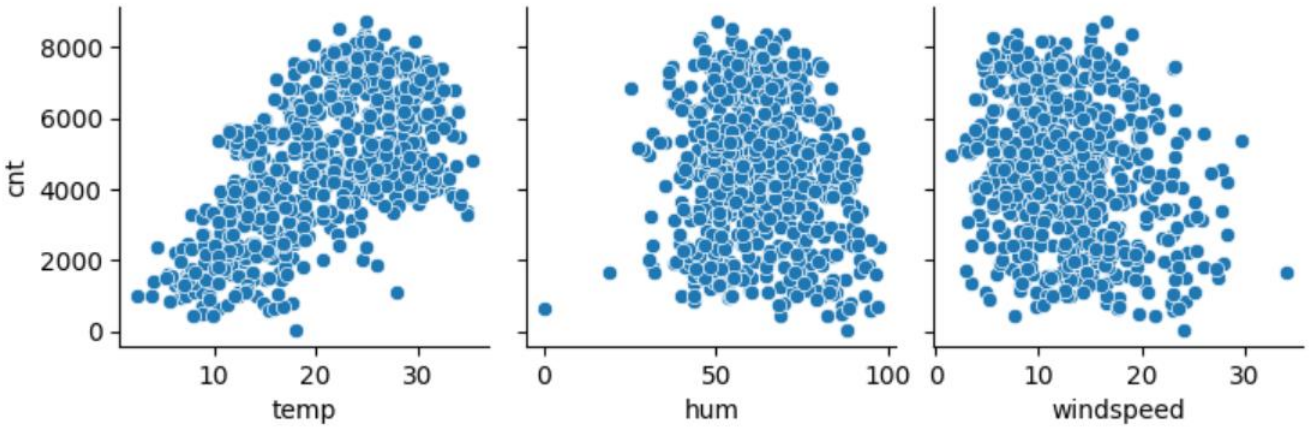
From these two plots we observe that :

- The season, year, month & weathersit have the most impact on the target variable. This can be observed from the barplot as well as the correlation values these variables have with the 'cnt'.

- Holiday, weekday and workingday don't have much impact on the target variable. All these have a absolute correlation value of around 0.06. This is also confirmed from the bar-plots of these variables.

Ans 2 : To understand the importance of drop_first=True during dummy variable creation, let us consider an example where we have a column named gender with two values male & female. Now, if we create dummy variables for gender with 'drop_first=False', we will get two dummy variables (one for male and another one for female). If we keep both of them, they will be having high correlation and at the same time it increases the number of columns to analyze. As this is not required, so we use 'drop_first=True' which drops one variable (say female). If male=1 then the person is a male and if male=0 then the person is female. Hence, 'drop_first=True' removes the redundant feature which increases the correlation as well.

Ans 3 : Lets look at the pair plots of the numeric variables against the target variable.



We make following obervations from the above plot :

- Variable hum & windspeed don't seem to have a linear relationship with cnt.
- Among the three variables i.e. temp, hum, windspeed it is clear that **temp has higher correlation with cnt** compared to the rest.

Ans 4 : Following are the assumptions and their validation :
1) **Linear Relationship** between the features and target: According to this assumption there is linear relationship between the features and target. Linear regression captures only the linear relationship. This was validated by plotting a scatter plot between the features and the target. For the given dataset, temp and cnt showed somewhat linear relationship.

2) **No Multicollinearity** between the features: Multicollinearity is a state of very high correlations among the independent variables. Pair plots and heatmaps (correlation matrix) were used for identifying highly correlated features. Variable 'atemp' was removed based on its correlation with 'temp'. Further at each iteration of the model building, vif values were checked and variables with high vif values (>10) were dropped.

3) **Normal distribution of error terms**: Another assumption is that the residuals should follow a normal distribution. After the model building, distribution of the residuals was checked with their 'distplot' and it showed that the residuals are normally distributed with a mean 0.

4) **Residuals are independent** of each other : Residuals should be independent of each other i.e. they should not exhibit any pattern. This was confirmed by the plot of residuals vs y_train_pred.

5) Error terms have **constant variance** (homoscedasticity) : This assumption was again validated by the plot of residuals against y_train_pred where we didn't see any visible change in variance of the residuals.

Ans 5: Based on the final model, following are the top 3 features contributing significantly towards explaining the demand of the shared bikes :

- **temp** : As was also evident from the plots & correlation, temp is the feature with maximum impact on shared bikes demand.
- **light-rain** : This is a dummy variable created from **season** and has a -ve coefficient. As expected as well, if it rains the bikes demand will go down. This behaviour is also seen from the bar-plot of season against cnt.
- **yr** : Since the bike-sharing systems are slowly gaining popularity, the demand for these bikes is increasing every year proving that the column 'yr' is a good variable for prediction.

Answers to "General Subjective Questions" :

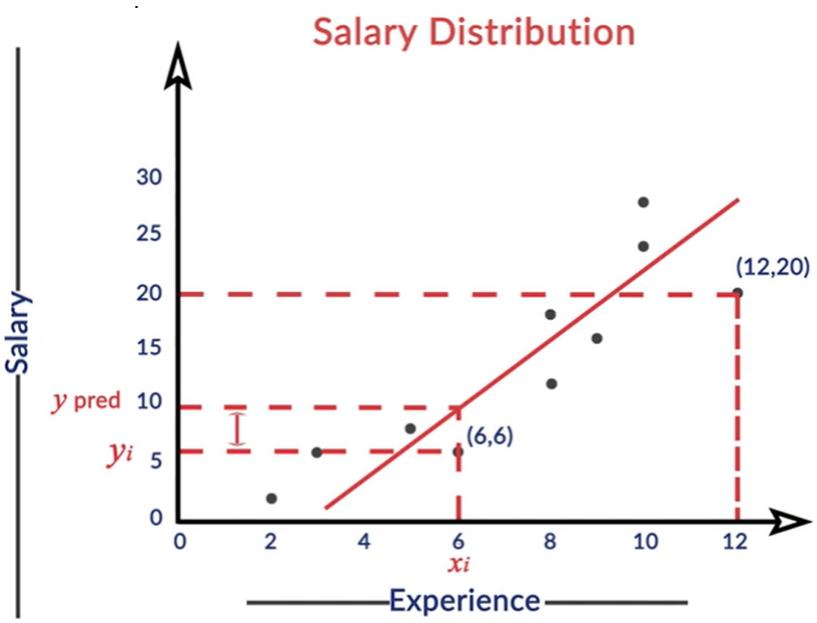
Ans 1 : The word regression means the act of going back. In the context of linear regression, it might mean that we need to go back to past values to predict a future value. The term "regression" was used by Francis Galton in his 1886 paper "Regression towards mediocrity in hereditary stature". He only used the term in the context of regression toward the mean. The term was later adopted by others to get more or less the meaning it has today as a general statistical method.

Linear Regression is used to establish a linear relationship between continuous variables such as the number of products sold, revenue of a company etc. To understand this better, lets consider an example . Lets say we want to predict the salary of a person based on the years of experience he has. We are given a sample data as shown below.

Sample Data

Years of experience	Salary (lakhs per annum)
6	6
12	20
3	6
8	12
9	16
10	24
2	2
5	8
8	18
10	28

Lets draw these points on a scatter plot. Now we have to fit a straight line that best represents the data points we have. The equation of the best fit line will be : $y_{pred} = mx+c$



At a value say xi, we will have a value yi which is coming from the dataset and the ypred lying on the line as shown below. Now, we would like to define a cost function which is based on the difference between yi & ypred for various points. In this example, for fitting a straight line the cost function will be the sum of squared errors i.e.

$J(m,c) = \sum (y_i - y_{pred})^2 = \sum (y_i - mx_i - c)^2 = (6 - 6m - c)^2 + (20 - 12m - c)^2 + \ldots$

To minimise the sum of squared errors and find the optimal 'm' and 'c', we need to differentiate J(m,c) w.r.t the parameters 'm' and 'c'. The resulting linear equations can be solved to obtain the required values of 'm' and 'c'. These values of m & c provide us the equation of the best fit line which can be used to predict the y values for a given x. Instead of doing these steps manually, we can use the python packages statsmodels & sklearn to build/predict the linear regression model.

Ans 2 : In 1973, a statistician Francis Anscombe came out with a quartet comprising of four data sets that have nearly identical simple descriptive statistics, yet look very different when graphed. They were constructed to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties. As shown below, each dataset consists of eleven (x,y) points (where the x values are the same for the first three datasets).

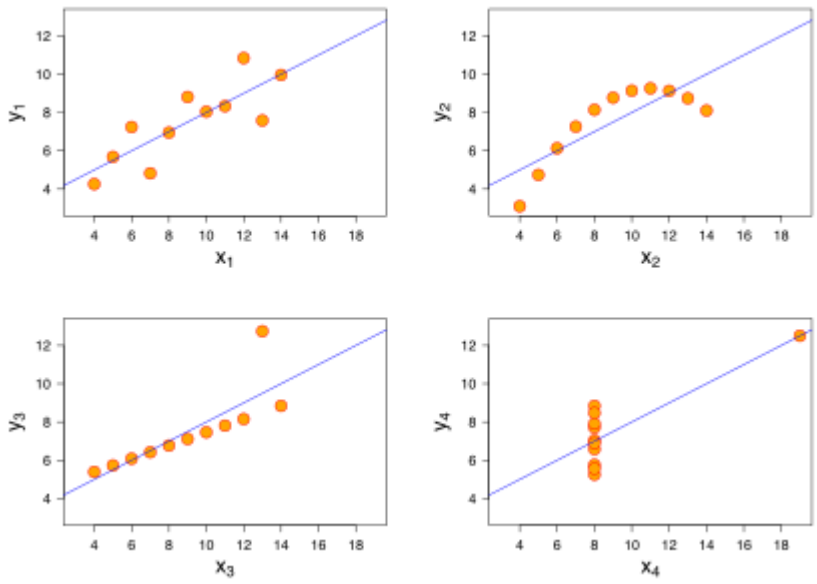
Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

As shown below, all the four data sets have nearly the identical statistical properties :

Property	Value
Mean of x	9
Sample variance of x : σ^2	11
Mean of y	7.50
Sample variance of y : σ^2	4.125
Correlation between x and y	0.816
Linear regression line	$y = 3.00 + 0.500x$
Coefficient of determination of the linear regression : R^2	0.67

Now, if we draw the plots for the four data sets, we see that even though the statistical properties of the four data



are quite similear, their graphs are very different. Now we see the real relationships in the datasets start to emerge. Dataset I consists of a set of points that appear to follow a rough linear relationship with some variance. Dataset II fits a neat curve but doesn't follow a linear relationship (maybe it's quadratic or parabolic). Dataset III looks like a tight linear relationship between x and y, except for one large outlier. Dataset IV looks like x remains constant, except for one outlier as well.

Computing summary statistics of the data wouldn't have told us any of these stories. Thus, Anscombe's quartet gives us an idea on how important it is to visualize the data to get a clear picture of what's going on.

Ans 3 : A correlation between variables indicates that as one variable changes in value, the other variable tends to change in a specific direction. Understanding that relationship is useful because we can use the value of one variable to predict the value of the other variable. For example, height and weight are correlated—as height increases, weight also tends to increase. Consequently, if we observe an individual who is unusually tall, we can predict that his weight is also above the average.

In statistics, a correlation coefficient is a quantitative assessment that measures both the direction and the strength of this tendency to vary together. There are different types of correlation that you can use for different kinds of data. The most common type of correlation is Pearson's correlation coefficient.

The Pearson correlation coefficient (usually just referred to as correlation coefficient) is the numerical correlation between a dependent (Y) and independent (X) variable. It results from analyzing the difference between X and Y – and the proposed mean. The overall equation to calculate the Pearson correlation coefficient is given as :

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Here, the numerator represents the deviation present in the sample with X and Y *together*, while the denominator represents the deviation in X and Y individually. These deviations give us an idea of how the deviation in X relates to the deviation in Y. There are different types of correlations :

- Positive correlation: Variable X and Y move in the same direction. For example, as one variable increases, the other variable increases too.
- Negative correlation: Variable X and Y move in opposite directions. For example, as one variable increases, the other variable decreases.
- No correlation: There is no apparent link between the two variables.

Ans 4 : Scaling is used to scale the data of an attribute so that it falls in a smaller range, such as -1 to 1 or 0 to 1. The goal of scaling is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. For machine learning, every dataset does not require scaling. It is required only when features have different ranges.

For example, consider a data set containing two features, age, and income. Whereas age ranges from 0–100, income ranges from 0–100000 and higher. Income is about 1,000 times larger than age. So, these two features are in very different ranges. When we do further analysis, like multivariate linear regression, for example, the attributed income will influence the result more due to its larger value. But this doesn't necessarily mean it is more important as a predictor. So we scale the data to bring all the variables to the same range.

There are several ways of scaling, the two important ones with their differences are as follows :

- Normalized scaling : This rescales the values into a range of [0,1]. This might be useful in some cases where all parameters need to have the same positive scale. However, the outliers from the data set are lost. Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. It is give as : $\mathbf{X' = (X - X_{min})/(X_{max} - X_{min})}$
- Standardized scaling : This rescales data to have a mean (μ) of 0 and standard deviation (σ) of 1 (unit variance). Standardization can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. It is given as : $\mathbf{X' = (X - \mu)/\sigma}$

Ans 5 : Variance Inflation Factor (VIF) calculates how well one independent variable is explained by all the other independent variables combined. Mathematically, it is given as : Vif = $1/(1 - R^2)$

Infinite value of Vif shows a perfect correlation between the independent variables. Lets consider a case of two independent variables. In the case of perfect correlation, we will have $R^2 = 1$, which will lead to $1/(1- R^2)$ infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

Ans 6: The q-q (quantile-quantile) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Quantile means the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below that value.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Apart from comparing the distributions of two data sets, the q-q plot can also be used to check the distribution of a sample data against a theoretical distribution. For this, q-q plots take your sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution. The number of quantiles is selected to match the size of your sample data. While Normal Q-Q Plots are the ones most often used in practice due to so many statistical methods assuming normality, Q-Q Plots can actually be created for any distribution.

Q-Q Plot in linear regression : This is used to assess if the residuals are normally distributed. Basically what we are looking for here is the data points closely following the straight line at a 45% angle upwards (left to right) as shown below :