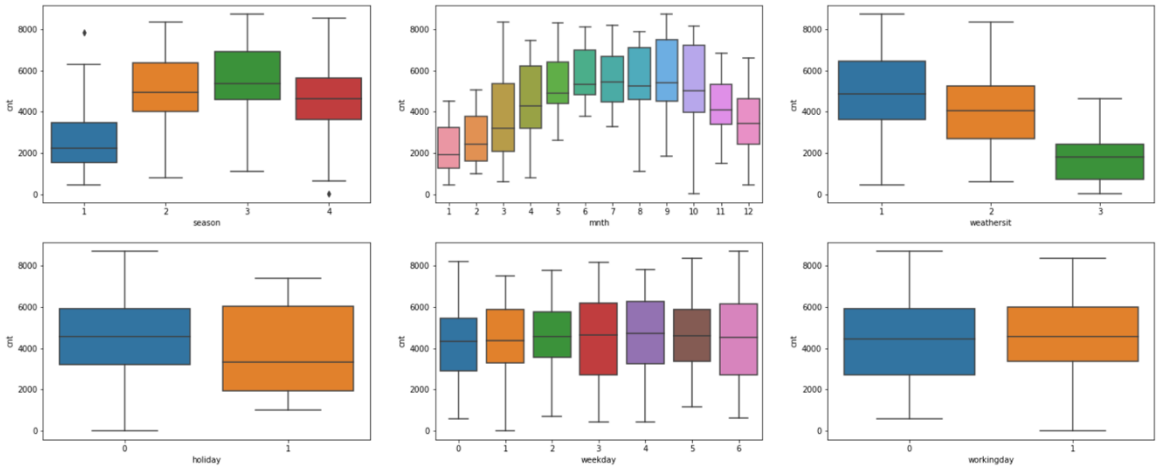
Assignment-based Subjective Questions

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

Answer:

- Season: Most of bike rental happen in season 3 with a median of over 5000 which can be a good predictor of our model
- Month: More bike rental happen in month 5,6,7,8,9 and with an average number of bike rental 4000 this can be a good predictor
- Weathersit: Most of the bike rental happen in weathersit 1
- Weekday Wednesday tend to have more than 10 % of the bike rental we will its effect on the dependent variable
- Holiday Over 90% of bike rental happen when there is no holiday
- Working day there is a slight increase on bike rental don working day



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

Answer: The **drop_first =True** parameter **pd.get_dummies(x, drop_first = True)** ensure that the first dummy variable is dropped and yielding n-1 dummy variables. Suppose we have a categorical variables whose values are 5 then we should get 4 dummy variables (5 - 1). This is important because as for 5 values can be represented by 4 keeping in mind that the remaining is absent and already encode in the 4.
Example: consider a categorical variable income with 3 values Low, Middle and High of we represent it using dummy we will get two variables such:

Dummy_middle	Dummy_high
1	0
0	0
0	1

In the absence of both middle and high we can infer that that this a low income household

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

Answer: The **(temp)** temperature and **(atemp)** feeling temperature have the highest correlation we can see from this plot that as the temperature increase there is also increase in number of bikes rented

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

Answer: After building the model on the training set, we validated out assumption on the test set.
Before doing that we made sure that the model coefficients have a small p values and small VIF.

1. Residual analysis: the distribution of errors terms should be normal with a mean of zero
2. We evaluated the model on the test set which the model has never seen before and R2_score was 0.80 closer to the r2 of the training set

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

Answer: Based on the final model the 3 top features are

1. Season Winter with a coefficient (season_4) of 0.6939
2. Season Summer with a coefficient (season_3) of 0.5574
3. Feeling temperature with a coefficient (atemp) of 0.4837

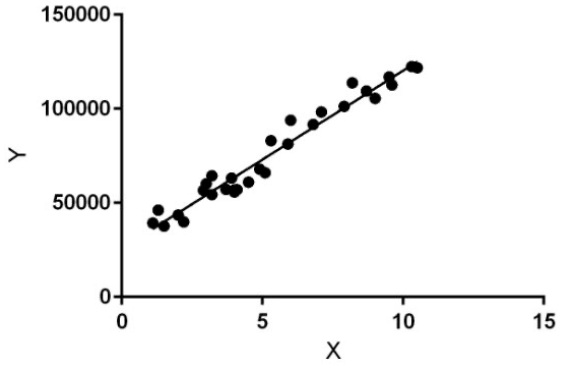
The P-values of these features are very low 0.000 and the VIF also.

General Subjective Questions

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

Answer: What is linear regression? Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.
As a running example, suppose that we wish to estimate the prices of houses (in dollars) based on their area (in square feet) and age (in years). To develop a model for predicting house prices, we need to get our hands on data consisting of sales, including the sales price, area, and age for each home. In the terminology of machine learning, the dataset is called a training dataset or training set, and each row (containing the data corresponding to one sale) is called an example (or data point, instance, sample). The thing we are trying to predict (price) is called a label (or target). The variables (age and area) upon which the predictions are based are called features (or covariates).

The line is the estimate of our fit.



At the heart of every solution is a model that describes how features can be transformed into an estimate of the target. The assumption of linearity means that the expected value of the target (price) can be expressed as a weighted sum of the features (area and age):

price = $w_{area} \cdot area + w_{age} \cdot age + b$.

Collecting all features into a vector $x \in R^d$ and all weights into a vector $w \in R^d$, we can

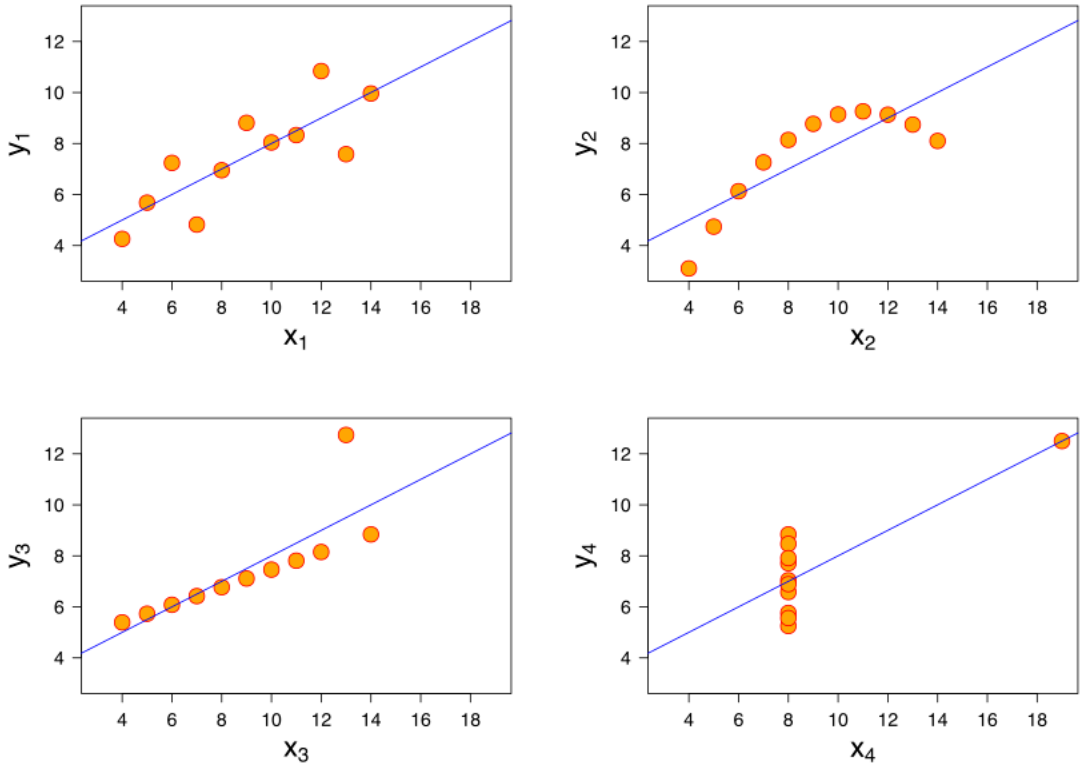
express our model compactly via the dot product between w and x:
$\hat{y} = w^T x + b$.

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

Answer: Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distribution and appear very different when graphed. Each dataset consists of eleven(x, y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and

the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough

---



- The first scatter plot (top left) appears to be a simple linear relation, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on *x*.
- The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

The datasets are as follows. 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

It is not known how Anscombe created his datasets. Since its publication, several methods to generate similar data sets with identical statistics and dissimilar graphics have been developed. One of these, the *Datasaurus Dozen*, consists of points tracing out the outline of a dinosaur, plus twelve other data sets that have the same summary statistics. *Datasaurus Dozen* was created by Justin Matejka and George Fitzmaurice. The process is described in their paper "Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing".

The Datasaurus Dozen, just like Anscombe's Quartet, shows why visualizing data is important, as the summary statistics can be the same, while the data distributions can be very different.

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

Answer: The Pearson correlation coefficient (*r*) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Pearson correlation coefficient (*r*)	**Correlation type**	**Interpretation**	**Example**
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is **no relationship** between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

The Pearson correlation coefficient (*r*) is the most widely used correlation coefficient and is known by many names:

- Pearson's *r*
- Bivariate correlation
- Pearson product-moment correlation coefficient (PPMCC)
- The correlation coefficient

The Pearson correlation coefficient is a descriptive statistics meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

Although interpretations of the relationship strength (also known as effect size) vary between disciplines, the table below gives general rules of thumb:

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

Answer: It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.It is important to note that **scaling just affects the coefficients** and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and
  1. **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

- **sklearn.preprocessing.scale** helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

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

(3 marks)

Answer: If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well)

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

(3 marks)

Answer: 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.

Few advantages:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior

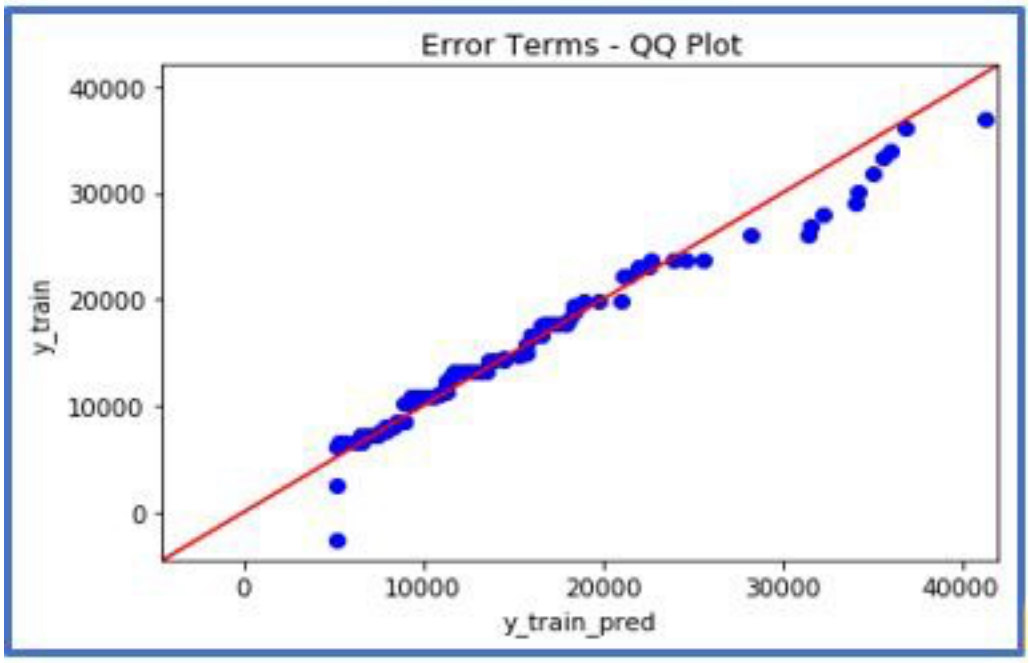
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

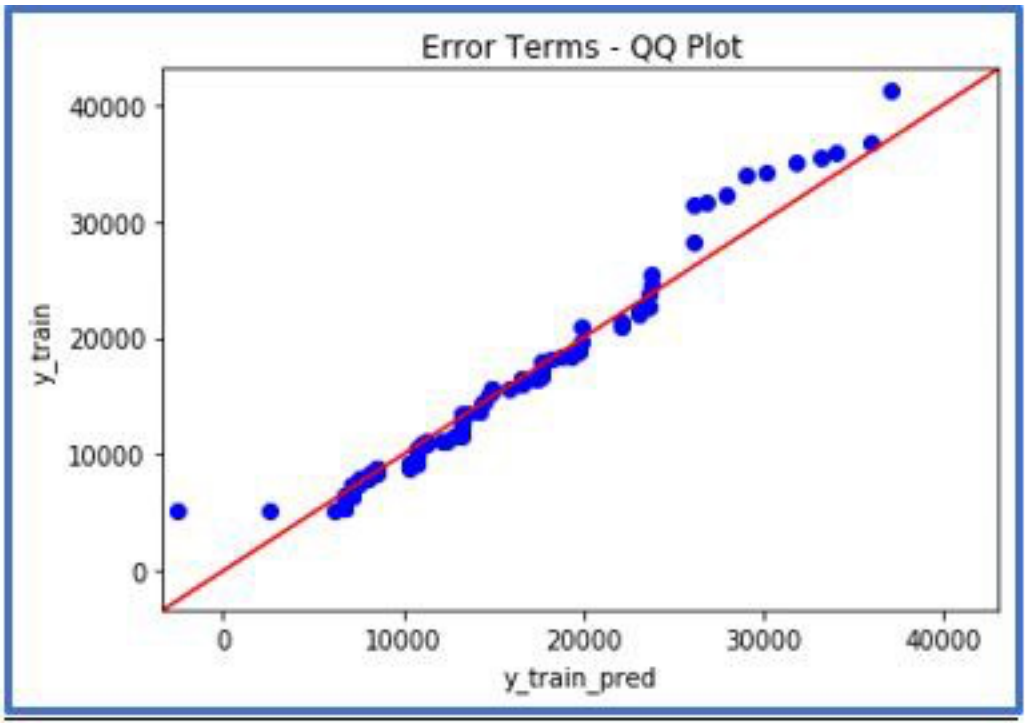
Below are the possible interpretations for two data sets.

a) **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) **Y-values < X-values:** If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: *If x-quantiles are lower than the y-quantiles.*



d) **Different distribution:** If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis

Python:

statsmodels.api provide **qqplot** and **qqplot_2samples** to plot Q-Q graph for single and two different data sets respectively.