**Theory questions assignment – data stream**

1. **What does ‘Data cleansing” mean? What are the best ways to practice this?**

* Data cleansing is the process for fixing or removing incorrect, duplicate or incomplete data within a dataset in order to ensure insights and analysis are useful and minimizing the risk of drawing wrong or inaccurate conclusions.
* In an ideal world we would want to keep as much as the data as possible when we are data cleansing in order for any analysis to have as much data behind it as possible.
* One of the first ways to practice this is looking at the NA values, if a variable has the majority missing values it may be best to remove it but if there are only a few then there a different ways of working with NA values. One common way would be finding the mean of a column and replacing all the NA values in the column with the mean. To improve this we could base it on further observations. For example if another variable is a category finding the mean of the column with NA values for each group in a category and replacing NA values with that.
* We would be to look if there are any outliers in the data that we may want to remove before we run the analysis and run a line of code to drop any duplicate rows

1. **What is the difference between data profiling and data mining**

* Data mining refers to the process of identifying patterns in a pre-built dataset in order to derive more useful bits of knowledge which can be fed into different areas of a business of instance
* Data profiling on the other hand looks at a dataset to determine the actual content, structure and quality, in order to do this it therefore analyses the raw data In other to collect statistics e.g: mean, median and std.
* The purpose of data mining is to mine the data for actionable information. Therefore you may apply computer-based algorithms to extract the information. Common tasks would be classification, regression and clustering.
* On the other hand the aim of data profiling and create a knowledge base of accurate data about the data, common tasks would be summary statistics mentioned earlier
* Data profiling on the other hand looks at a dataset to determine the actual content, structure and quality, in order to do this it therefore analyses the raw data In other to collect statistics e.g: mean, median and std.

1. **Define Outlier with an example**

* An outlier is an observation that lies an abnormal distance from other values in a random sample in a popular. Outliers could be errors or instances we want to exclude but other times we may want to include them
* There are different ways to observe outliers once way is to look at the interquartile range and calculate:
  + Q1 -1.5IQ and Q3 + 1.5IQ, any value that are outside of this could be determined to be outliers in a dataset
  + Data outside 2 standard deviations from the mean can also be considered outliers
  + [1,2,3,4,5,6,7,8,9,100] we could consider 100 to be an outlier as if we look at the below python code both methods 100 is outside the outlier range (outlier1 and outlier3)



1. **What is collaborative filtering?**

* Collaborative filtering filters are a family of algorithms, which by using interactions and data collected by the system from the user and other users can recommend items.
* For a user Collaborative filtering only is based on historical preferences and the core assumption is that users who have agreed in the past tend to also agree in the future
* There are two ways of preference: explicit: is a rate given by a user to an item, whereas implicit is indirectly such as page views.
* Collaborative filtering will focus on the relationship between users and items and there can be two classes:
  + User based – which measures the similarity between target users and other users
  + Item-based which measures the similarity between items that target users rate or interact with and other items

1. **What is time series analysis?**

* Time series analysis is a way of analysing a times series of data. A time series of data is a sequence of data where you have collected data points over an interval time at consistent intervals instead of doing it random. For example points can be collected weekly, monthly or yearly.
* The data would be ordered from oldest to newest data and the purpose of analysing this is to see how variables change over time whereas non-time series data will not show that.
* Time series analysis accounts for the fact these data points may have an internal structure for example follow a trend or have a season variation which should be observed
* Applications of time series analysis include for forecasting
* An example of time series analysis would be looking at seasonality, which is looking at predictable changes that occur over a one-year period.

1. **Explain the core steps of a Data Analysis project?**

* The first step is to define the question you are trying to answer through the project this way you can have a clear objective and can measure if the project has been a success
* The second step is to obtain the data you will need to answer the question this may be by querying a database in SQL, through connecting to an API or downloading data
* The third step is to clean the data to try and increase the chance of quality outputs from the analysis. To do this you will need to remove/replace any NA values and remove inaccurate data / outliers
* The next step is to explore the data to understand it better through looking at different statistics and visualising the data through charts. By exploring the data you can start to find different trends
* The next step would be to do any modelling on the data this could be for predicting for example using linear regression or to understand the data better through clustering. After doing this you would want to assess how well your model works, through statistics eg Rsquared or through a confusion matrix. If the model doesn’t work you would want to iterate and perhaps remove some of the variables that don’t have a high correlation or change aspects of the model.
* The last step would be to write out a report to explain your findings and communicate this with other people

1. **What are the characteristics of a good data model?**

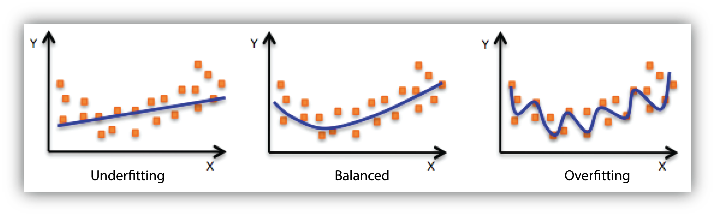
* A good data has a predictable performance meaning we trust it to perform consistently
* Can adapt to changes in requirements while still predictable performing
* The model has quality outputs and will be able to produce outputs that are useful and answer the question we set out
* Doesn’t have parameters which do not add any predictive power / are useful for the model

1. **Explain and provide examples of univariate, bivariate, and multivariate analysis?**

* Univariate analysis is the analysis of one variable at a time and therefore is the simplest form of data analysis, as it doesn’t look at any causes or relationships between variables. However it does look at describing the data and looking at any patterns, this could be done by looking at the mean, variance, std etc. An example would be looking at a sample of different heights.
* Bivariate analysis looks at analysing two variables for example looking at height and weight and therefore we can see if there is a relationship between the two different variables by looking at the correlation and plotting it on a scatter plot.
* Multivariate analysis looks at analysing more than two variables for example looking at height, weight and age. Due to having more than 2 variables there will be lots of different ways to analyse the relationships between the variables one way would be through linear regression.

1. **What is linear regression?**

* Linear regression is a type of predictive analysis looking at a set of independent variables (x), which will predict the dependent variable (y). For each independent variable there will be a coefficient, which signifies the relationship between that independent variable and the dependent variable.
* The formula for linear regression in the simplest form is y = c + b\*x where y is the dependent variable, c is a constant and b is the coefficient on the independent variable. The formula is one for a line, hence the predictions will be linear.
* In order for linear regression to model a relationship there are assumptions that need to be made which are the following: 1) the relationship is linear 2) errors are normally distributed 3)Homoscedasticity of errors (the variance of the error terms are constant) 4) Independence of the observations
* There is a range of uses of linear regression, which include forecasting, or finding the cause and effect relationship between variables.
* In order to evaluate how good a linear regression model is we can use R-Squared – the closer to 1 the better the model is, however if it is really close to 1 the model may be overfitted.

1. **In terms of modelling data, what do we mean by over-fitting and under-fitting**

* Using the diagram above, underfitting is when the model performs badly on the training data and test data, as the model hasn’t captured the relationship between variables and structure of the data. In the picture above we can see the model has failed to pick up the slight upticks on the left and right.
* Over-fitting is when the model performs well on the training data but poorly on the test data. This is because the analysis corresponds too closely to the training data but therefore will fail to fit additional data – in this case the model behind to memorize the training data instead of learning the relationships and trends in the data. In the chart above we can see the line has taken in account nearly every data point even those which don’t follow the general trend.