# 2023463 - MSc in Data Analytics

Author: Umar Iqbal

e-mail: [2023463@student.cct.ie](mailto:2023463@student.cct.ie)

Student ID: 2023463

**Data Preparation and Visualization:**

Our journey started with selecting the dataset for our project. As we have been asked to select it from CSO website so we chose the data on monthly unemployment. The dataset contains following features:

It has **Statistic Label** which is continues variable and contains the same value in it. It tells what this dataset is about.

Second Column is **Month,** which is type of object and contains year and month in it, like this 1998 April.

Third Column is **Age Group,** which is a categorical column as well and contains different age groups in it like in this format ’14 – 24 years’.

Fourth Column is **Sex,** which is of type object and contains genders in it.

Sixth Column is **Unit,** this column holds’%’ as value in it and tells that the next column’s value is in percentage.

Seventh Colum is **Value In Percentage,** It hold unemployment rate in percentage.

Eight Column is **Unit** again and it holds ‘Thousands’ as value in it. And tells us that the next column holds value in thousands in it.

Ninth Column Is ‘VALUE In Thousand’ which holds unemployment rate in thousands.

We will be using this dataset for our work. So, it started with exploring with the dataset. A good approach to tell that what stories we tell with data. In the Data Preparation and Visualization, we just did not focus on getting insights but we prepared the data for Machine learning models as well.

Well to begin with the project we first import the required libraries. These libraries are our working hands and we use methods of these libraries to get our work done.

Then we loaded our dataset into notebook so we can work with it. We used pandas(library) for this. Pandas helped us to get the data into the notebook to work with it. Then to check that what we have inside our dataset and is it loaded into our notebook we used a method called **head ()** to look into our data. It displays first few observations of our data. Then we used **tail ()** to view the last few observations of our dataset. With the help of this we sneak peek into the dataset.

We started working with our dataset but we need to know t better in dept. So we dived into the dataset and used **shape ()** method to know the total number of columns and observations in dataset and it returned us that we have 1854 rows and 8 columns in our dataset. Then we looked for column names in our dataset and we used column keywords to get just the column names in our dataset.

Then we checked for missing values and we know that missing values are gaps in our memory and if we have missing data so our machine learning models can’t predict well. So we deal with such issues like missing values in our data preparation step. If we get missing values, we fill them with mean/median. But Luckly we don’t have any missing values in out dataset. We are checking for missing values using isnull() function and them asking our notebook to sum all the missing values of each column and show it to us.

But it returned zero. And we also visualized it using heatmap which showed us that we don’t have any missing values in our data. So now completeness of our dataset is confirmed. After that we looked for any duplicate values in our dataset. Looking for duplicate is also important but it can cause of getting wrong results. We used the same way to remove the duplicates. We used **duplicate ()** method and asked to sum up using **sum ()** function so we can see that how many duplicates each column contains and fortunately we have no duplicates in our dataset.

The next step was to look for data types of our columns. So, we used **info ()** method to check for data types. And here we got to know that every column has correct datatype, e.g., value in thousands and value in percentage has datatype of float and age group, sex statistic label, unit has object type which is okay. But the month has object type as well. Since it is a data so converted it to date format like this “1998-01-01”.

Then we used described method to know more about our data like what is the middle value, what is extreme. Then I used boxplots to detect outliers in our dataset. Boxplot gave us glimpse in variability of 2 columns named “Value in Percentage and Value in Thousands”. Then we visualized different categorical columns like gender and age group. With the help of these graphs, we know that how our data is divided into different categories. Then I looked for data consistency. I displayed the first and last few rows of the dataset and found that it is consistent.

So now we know about out data so let’s play with it. Let’s get it ready for machine learning. So, we stated our work with the first step of tuning our data to work it with machine learning models.

As we know that humans can read data in different formats but computers can’t. Machine prefer consistency. So, we extracted the year and month from this column and created the 2 features from this column (Month). This way our data makes more sense. Years are stored in year column and months are stored in Month Number. Machine learning models work better with numbers instead of categorical values so we converted our categorical columns like age group and sex into numbered columns. And we removed both of the columns ‘UNIT’. Because they are not of any use. Now our data is clean and simple. If you look into it. It got so simple.

After doing feature engineering our data makes more sense. We can understand it by this example that after feature engineering we got a lot of numbers but with these numbers we get to know about people work life like their age and how many of them are working and their genders and the time period as well.

Visualizations are way to display things easily. So, if a human sees a graph he understands what is going on there. We used histograms to get sense of distribution of our unemployment rate. And the line on that distribution gave us visual of underlying probability distribution.

Then we showed the line plot of unemployment rate for both sex in different age groups. The line plots are easy way to show the change over time. And I used different colours for different age groups. And after that I used boxplot to display the spread of data. I showed unemployment rate for different age groups and genders side by side. This makes picture clearer that in which age group which gender has more unemployment rate. Every graph I used is appropriate and there are different colours and legends added to understand it better. Also, everything in code is also explained well.

**Statistics for Data Analytics:**

In Data preparation step we made our data ready for further process and we did feature engineering of our data as well. The dataset got very detailed and each row of it tells a story like how many people are unemployed and what is their age and gender. And there is not missing row in the data as well which is good. Now we want to solve this puzzle more by knowing the full picture about unemployment rate. Now we have a lot of numeric data in our dataset so we will use some math and work with our data to have more understanding of unemployment rate.

So, I gigged into the data more to perform the Descriptive Analysis. We looked for average unemployment rate which is called mean, and middle value which is not too high and not too low which is median, also how spread out the numbers are means how much numbers jump from the average and this is called variance and standard deviation in descriptive analysis. We used built in functions to perform this. Those functions are mean(), median(), var(), std().

First of all I have calculated mean, median, var and std for the whole dataset. We calculated it using the Value In Percentage and Value In Thousands columns. As we know that our data is grouped. It has different age groups and different genders so I dived more into the data and performed critical analysis like. I calculated these measures for each gender and each age group. By doing this we got more sense of our data. So I calculated all these measures for Male and then for female and then for one age group and then for other. With help of this we tried to know about our dataset more just like when we meet someone so we ask questions to them get to know them.

After that we displayed some graphs to talk about our data. First of all, we used line chart to display the yearly unemployment rate. Line graphs are best to show the data over time. Because it gets very clear to see where it goes up and where it comes down. With the help of this we clearly can see that what is year where unemployment rate is high and which year has lowest unemployment rate.

Then I created Boxplot to look in detail about unemployment rate. I used Value In Percentage and plotted 2 boxplot graphs to check for unemployment rate. One graph is showing the unemployment rate by Age groups and other by gender. By doing this we can easily see that who is getting hard time in job market. We can see if males have higher unemployment rate of females. Or we can check if specific age group has higher unemployment rates.

Then we moved to the next step which is to apply discreet distribution on our dataset. We have been given two options here, first is binomial distribution and second is poisson distribution. It is like making a good guess. Let’s understand it detail and see what we used on our dataset.

Binomial distribution is used when we have yes or no questions just like a coin which has 2 sides. If we flip a coin, we only get head or tail. In such cases the binomial distribution is best to use. And on other hand poisson distribution is counting something for a number of times over a period of time. It is not like yes and no; it is like how many times this happened.

I chose poisson distribution here. It is good because we are checking the chances of something happening, for example how often young people might not be able to find the job. I tried to dig deep into it and created multiple charts for it like for each month. Because there could be time when it could be hard to find the job. Some months can have more unemployment and maybe some not. Well with the help of this we are not just trying to print some numbers on screen we are trying to figure out a pattern so we can learn from it and get prepared. We can get to know that when young people are struggling more to find jobs.

Then we checked if our data is in normal distribution. Suppose we have a room of students and we check for average height of students so the most of the students will be of middle height and few will be tall and some will be short. So, we checked for the same in our unemployment dataset. We used Q-Q plot to check if data is in normal distribution but it wasn’t. so, then we applied the log transformation to get our data into the normal distribution. So, numbers were all over the place. Some numbers were very big and some were very tiny. So, it doesn’t look smooth. When we apply log transformation it presses the big numbers to make them small and stretch the small numbers to make them big. It just gets the numbers together and it never changes the order or numbers. So, log transform makes the data uniform. So, after applying the log transform, we got our data into bell shaped curved. I also displayed the data distribution after applying the log transform and then displayed the Q-Q plot to see if it follows the straight line. And it almost followed the straight line. So, we can say we got most of our data into normal distribution.

**Machine Learning for Data Analytics:**

As we know that we have made or data ready for the final step of our work. We performed EDA, we cleaned our data and we did feature engineering to make it ready for machine learning and also, we applied normal distribution on our data too.

Learning The next step if of selecting features which is very critical steps. We have to analyse our dataset critically to choose variables Selection of correct features is important because it affects the result. So we carefully chose our variables that we thought will be good for our analysis for example we chose age group,, gender, year and month. Because they have relation with each other. The unemployment rat depends on different things.

After selecting features, we split our data into training and testing. We don’t change the feature selection for training and testing. The feature selection remains same. So, we practice on our data we make sure that everything is perfect by practicing before getting on any conclusion. So, this way models learns with the training data and then use the testing data to check if it can apply what he learnt from the training. This step is good and important because we can see that if our model is good at predicting.

Then we applied 2 regression models, Linear Regression and Support Vector Regression. First of all, we used the linear regression on our data which is the simplest. Regression. It shows us that how different things like age group and time of year affect the unemployment rate. It helped us make sense of different things in very simple and easy way.

Linear Regression is just like a straight road. It is easy but as we know that things get complex sometimes. I tried another regression model on our dataset to see how that one works. And this time I chose the complex one. Which is Support Vector Regressor. If linear regression is like a straight road so Support Vector is like curvy road or we can say wavy path. It helps us to get the story in more detailed way.

After that we have applied classification on our data. Classification is like sorting our data. It is just like how we organize things in real life. At word we keep or old files at a separate place. Our computer or laptop has their place to be on. So, when we keep things on their place, we actually organize them or classify them. We do the same with our data too. We classify the information into categories.

We have used two classifications in our notebook. First one is Random Forest And second is support vector classifier. Both of them did a very good job with one hundred and ninety nine percent of accuracy.

Next step was to check that how effective our Regression models performed, we used MSE which is to measure the average error of our predictions. We also calculated the MAE to check the size of errors and one more important thing which R2. R2 is use to understand that how good our model works with variation in our data.

I displayed this information of both regression models in a table. Which is easy and good way to understand the performances of different models. It is like we can easily do the comparison that which model performed best in the data.

As we know that we have applied different supervised learning model. So, the CRISP-DM was the good choice here. Let’s understand it with a simple example of making food. Suppose if we are making a special dish so we will not just put everything in it that we have in the kitchen. We have to know everything like what we are cooking and what ingredients we need for that. CRIPS is like the recipe for making sense of data. Let know we it guides us with the things. First of al we need to know the data; we need to know what problem we are going to work on. What we will be solving with our data. Then we get our data cleaned up for it. Clean dirty data, duplicates and fix missing values. So, this step is preparing the data. We are preparing it to solve our problem we identify in the first step. Then the modelling comes. This is the main step. We use our data that we prepped in previous step and we start playing with it. We apply different models. And we see what happens. After that we evaluate our data, the results we got we check them. It this important too. Like as we were talking about preparing a dish. So, suppose we prepared a dish, before serving it to guests we will taste it first we will check if it is perfect. This is what the evaluation means. So we evaluate our results. We used supervised learning in our assignment because we had the labelled data. Unsupervised learning is used when data is not with predefined labels. I also put the comparison tables after each model so anyone can compare the progress of all these models that which one performed well with the data.

**Python Programming:**

Well, we sued python programming to perform detailed analysis on our unemployment data. We will be talking that what programming techniques we used in this analysis and did we critically analyse each module or not.

As we know that python has a lot of libraries and they offer a very good platform for data analysis. In our Project we used many libraries like pandas for data manipulation, seaborn and matplotlib to visualise our data, scikit-learn for modelling and to do predictions and many more libraries.

Let talk about few libraries to understand them better, As I mentioned above that we used panda for data manipulation. Pandas handle data very smoothly. With the use of panda’s library, we can easily word wit tabular data. For example, with the use of different functions we can perform complex task so easily. Like we can use read\_csv to load dataset into our notebook. We can see how strong these libraries are that with the help of just one lien we can grab all the columns and rows and perform operations on that data. Then we use different more methods of panda library to work with our dataset like we used head () and tail () to view first and last few rows of our dataset. So, we looked at the both ends of our dataset with these methods. So, we can see that how this library helped us to process the data so easily and smoothly.

We also did data cleaning of our data using pandas. We used isnull () to identify any null values in our dataset and used duplicate to check for any duplicate values. We also looked for shape of our data and all the column names and their data types. So, we can see that how powerful this library is for data cleaning process.

As we know that visualization of our data set is also important. For that we used matplotlib and seaborn to create different graphs. For example we can see in our notebook hat we generated a plot by plt.plot() to show graph of unemployment rates over time. Graphs are easily understandable than numbers by seeing a picture. We crated multiple plots to understand the data more in depth.

When you see the notebook, you will see that the code quality through the end of cell is maintained. The notebook is readable and consistent. Every process is in sequence and explained after the result of it. Also, used graphs to make it more clear. Used comments to describe what I did in code. Code is indented properly and chose perfect function and variable names. I used camel notation to name my variables. You will see every variable of function I defined in code is in the same format of camel notation. If we maintain the code quality standards so this way our notebook gets easy to understand and debug and maintenance gets easy as well for future.

While doing our work we used different programming paradigms. You can see that the whole python notebook is very straightforward. Everything is written step by step. Every task is completed in an order from getting our data to notebook to perform predictions. It is just like giving instructions to go to from a point to another. For example, in cleanup process we looked for missing data first. It wasn’t like that to tell your notebook to fill the missing data first and then look if we have any missing data. This is Imperative programming paradigm.

We also used the Declarative Programming Paradigm in our notebook. As we can see in our notebook that we filtered our data like this “monthlyUnemploymentDF[monthlyUnemploymentDF['Age Group'] == '15-24']”.

It is just like SQL (structure query language) which is used to play with databases. In declarative programming we see what we want to do rather than how to do it. Pandas helped us a lot here by filtering and transforming the dataset.

Object Oriented Programming helped us in complexity. We created different objects of different classes like StandardScaler and LinearRegression to work with them. Creating objects help us using different methods of those classes. Like after creating the LinearRegression Object we can use different methods of that class like fit and predict to perform complex operation on our data.

We also used functional programming in our notebook. We used map () to change our data form string to numbers. It converted unemployment rate form string to numbers. Also, while working for statistics part, we defined a function which helped us to know that what will happen when the data size increases. So using functions to do our work makes life easy because they are reuseable and code gets readable.

We almost used all the programming paradigm which we studied so far. All these different ways make our work easy. We made the flow in sequence. By using all these our notebook got easy to read and got into a useful insight about our dataset.

**GitHub:**

<https://github.com/UmarIqbal3347/MSC_DA_InterGr_CA1.git>

here is the GitHub repository I have set up while working on this project. This is the public repository. So anyone with this link can see all of the files inside in this repository.

**References:**

For EDA I have read the “What is Exploratory Data Analysis.pdf” document that was given by professor in class. This helped me to get detailed idea about EDA and helped in the process.

For visualizations I read this “What Visualization should I use.pptx” power point. It was really very helpful understanding which visualization can be used for different scenarios.

For feature engineering I got a lot of help from “Encoding Categorical Data.pdf” and “Feature Engineering.pdf”. This really helped me because I learnt on hot encoding from these pdfs and applied in my assignment.

For statistics I used the class notes “SDA\_Lec 2.pdf” as well. The normal distribution and the poisson distribution I applied, is learnt from the class notes. Since I had the grouped data so I learnt from the class that how to calculate our measure in descriptive analysis.

For Machine Learning I also got help from the notes as well. And form the class practices which were explained in detail. Apart from these I read articles from internet too to get more understanding of things I did in this assignment.