THEORY QUESTIONS ASSIGNMENT

Data Science Stream

Maximum score: 100

KEY NOTES

- This assignment to be completed at student's own pace and submitted before given deadline.
- There are 10 questions in total and each question is marked on a scale 1 to 10. The maximum possible grade for this assignment is 100 points.
- Students are welcome to use any online or written resources to answer these questions.
- The answers need to be explained clearly and illustrated with relevant examples where necessary. Your examples can include code snippets, diagrams or any other evidence-based representation of your answer.

Theory questions	**10 point each**

1. **What does "Data Cleansing" mean? What are the best ways to practice this?**

   Data cleaning also referred to as data cleaning or data scrubbing is a complex set of tasks, it encompasses several processes aimed at improving data quality, the ultimate goal of data cleaning is to make a dataset consistency, reliable, valuable, and accurate as possible. There are many tools and practices to debug a data set. These processes are used to correct or delete inaccurate records in a database or data set. Generally, this means standardizing, identifying, fixing replacing, changing, updating, removing syntax errors, empty fields, incomplete, inaccurate, incorrect, incorrectly formatted, corrupted, duplicated in a dataset or records sometimes this occurs when we combine multiple data sources. This processes can vary from dataset to dataset because there isn't an absolute way to do it but a template for a process is important. Without the cleanup, the scan results are likely to be skewed.

   Data cleansing is an essential and key process of the data management process also is a relevant step of data preparation for further use whether in operational processes. After this process the data is ready for use in business intelligence and data science applications. The data cleansing is done by data quality analysts and engineers and data management professionals also the data scientists and business analysts may also clean data or take part in the data cleansing process for their own applications.

   The steps in the data cleansing process depends on the data set and analytics requirements but the usual steps are: **Audit: Inspection and profiling**. First, data is inspected and audited to identify issues and evaluate quality level this involves analyze the content of the data(data profiling) , documents relationships between data elements, verify data quality on data sets to find errors, discrepancies and other problems. **Definition of the transformation flow and mapping rules:** depending on the number of data sources, their heterogeneity and the anticipation of data quality problems, it will be necessary to execute more or fewer steps in the transformation and adequacy stage. The most appropriate is to propose an action at two levels, one at an early stage, which corrects the problems related to data coming from a single source and prepares them for a good integration; and another,

which intervenes later, dealing with the problems of data from a variety of sources. To improve control over these procedures, it is convenient to define the processes by fitting them into the specific framework. **Verification and validation**, after the cleaning step is completed, the person or team that did the work should inspect the data again to verify its cleanliness and make sure it conforms to internal data quality rules and standards. The level of adequacy and the effectiveness of a transformation action must always be tested and evaluated. Typically, this validation is applied through multiple iterations of the analysis, design, and verification steps; since some errors only become evident after applying a certain number of transformations to the data. **Reporting.** The results of the data cleansing step are reported to show up the trends and progress. The report also explain the number of found issues, the corrected data and metrics. The data that has been cleaned moves to the other steps of data analytics.

The data cleansing is important because it can boost your organization's competitive advantage the quality of data cleansing helps provide complete and reliable information to identify changing customer needs and stay on top of emerging trends. Data Cleansing can produce faster response rates, generate quality leads, and improve the customer experience. Data cleansing also increment productivity, where everyone can read and understand it because it always set to the same pattern, better prediction and modeling, saving time when working with it. False conclusions because of incorrect or “dirty” data can inform poor business strategy and decision-making.

2. What is the difference between data profiling and data mining?

Data Profiling	**Data Mining**
Data profiling is the act of analyzing the content of your data. Data profiling is an evaluation, that uses business tools and analytical algorithms to discover, understand, and potentially expose inconsistencies. This knowledge is then used to improve data quality as an important part of monitoring and improving data health.	Data mining is the tool that brings together different technologies and techniques capable of extracting such valuable knowledge from large databases. It is the process of collecting significant information which is explored and classified to identify patterns and establish relationships with which to make predictions and find explanations for the behavior of said data.
It can be executed on structured and unstructured data.	It is only executed on structured data.
It involves the discovery and analytical techniques to collect useful information related to the data.	It involves various techniques to perform tasks. Techniques like classification, clustering, regression, association rule and neural network.
A data profiling tool allows different types of analysis to be carried out among them are: ▪Completeness analysis: by looking at its results you will discover how often a given attribute is filled in, and how often it remains blank or appears as null. ▪Value distribution analysis: allows you to find out what is the distribution of records through different values for a given attribute. ▪Uniqueness analysis: it is the fastest way to know how many unique (distinct) values are found for a given attribute in all the records. Through this analysis you will easily identify duplicities. ▪Pattern analysis: it is the means through which data profiling makes it possible to know what formats were found for a given attribute and what is the distribution of records through that or other formats. ▪Range analysis: it is used to discover what are the minimum, maximum and average values that occur for a given attribute.	Some of the common techniques of data mining are association learning, clustering, classification, prediction, sequential patterns, regression and more. •Association learning is the most commonly used technique where relationships between items are used to identify patterns. It is also called relation technique. •Classification technique classifies items or variables in a data set into predefined groups or classes. It uses linear programming, statistics, decision trees, and artificial neural networks in data mining. •Clustering technique creates meaningful object clusters that share the same characteristics. Unlike classification that puts objects into predefined classes, clustering puts objects in classes that are defined by it. •Prediction technique predicts the relationship that exists between independent and dependent variables as well as independent variables alone. •Sequential patterns technique is used to identify similar trends, patterns, and events in it over a period of time.
The benefits of data profiling are: ▪ Poor data quality management can have negative effects on business operations and cost money.	The benefits of data mining are: ▪ Data mining uncovers information that was not expected. As many different models are used,

Organizations must also spend time rethinking strategies and rebuilding their reputations that's why companies must profile and monitor their inbound metrics. ▪ Improved data quality and reliability; by creating data profiles, organizations can ensure there are no duplications, null values, or anomalies. It also helps filter data, ensuring that the brand has useful and valuable information at hand. The quality and credibility of their data is important to make important business decisions. ▪ Make forecasts based on data; organizations can identify possible future outcomes relative to the market and their business and make predictive decisions with profiled information. This prepares the brand to address problems before they occur and allows them to effectively safeguard their financial health. ▪ Improved data organization, organizational data can come from a variety of sources, from business software to social media. Data profiling tools allow business teams to trace their metrics back to their source and ensure encryption for security.	some unexpected results tend to appear. The combinations of different techniques provide unexpected effects that become added value to the company. ▪ Huge databases can be analyzed using data mining technology. ▪ The results are easy to understand: people without previous knowledge in computer engineering can interpret the results with their own ideas. ▪ Contributes to tactical and strategic decision making to detect key information ▪ It allows to find, attract and retain customers and reduce the risk of losing customers. ▪ Improves the relationship with the customer: the company can improve customer service based on the information obtained. ▪ It allows you to offer your customers the products or services they need. ▪ The models are reliable. The models are tested and verified using statistical techniques before being used, so that the predictions obtained are reliable and valid. ▪ The models are generated and built quickly. Modeling sometimes becomes easier since many algorithms have been previously tested. ▪ It opens up new business opportunities and saves costs for the company.
The challenges of data profiling are: ▪System performance, data profiling involves large amount of column comparisons within, between, and across tables this requires a huge number of computational resources (memory and disk space), also more time to complete and build output results. ▪Limiting scope of results, since data profile reports are generated by summarizing and aggregating data values, there must be a threshold that defines the level of summarization to be implemented. The ability to limit or filter what goes in and what does not into the final profile report is a challenge. ▪Is important use profiled reports, generate data profiles, analyze datasets to understand the structure and content formation. ▪Self-service data profiling tools, the unavailability of self-service data profiling software tools is a common challenge faced. A self-service data profiling tool that can output quick 360-view of data and identify basic anomalies, such as blank values, field data types, recurring patterns, and other descriptive statistics is a basic requirement for any data-driven initiative.	The challenges of data mining are: ▪ Social and security challenges: decision-making strategies are made through the exchange of data collection. Sensitive information about individual is collected for customer profiles and understanding of behavior pattern. Illegal access to information and the confidential nature of information are becoming a major issue. ▪ User interface: The knowledge discovered through data mining tools is only useful if it is interesting and above all understandable to the user. ▪ Challenges of the mining methodology like the versatility of mining approaches, diversity of available data, dimensionality of the domain, control and handling of noise in the data, etc. ▪ Most data sets contain exceptions, invalid or incomplete information leads to complications in the analysis process, and some cases compromise the accuracy of the results. ▪ Complex data: the data Real-world data are heterogeneous and could be multimedia data containing images, audio and video, complex data, temporal data, spatial data, time series, natural language text, etc. It is difficult to manage these various types of data and extract the required information. ▪ The performance of the data mining system depends on the efficiency of the algorithms and techniques that are used. The algorithm must be efficient and scalable to extract information from large amounts of data in the database.

3. Define Outlier with an example.

An outlier is an abnormal and extreme observation in a statistical sample or time series of data that can potentially affect the estimation of its parameters. In simpler words, an outlier would be an observation within a sample or a time series of data that is not consistent with the others.

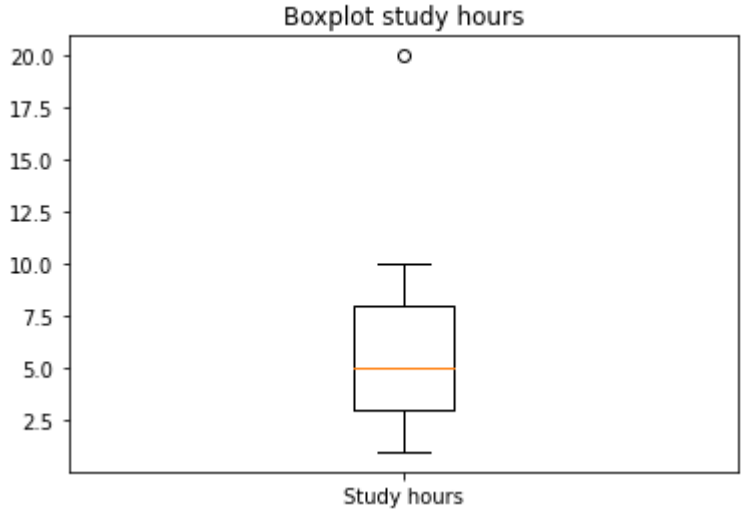
Example:

The next data are the hours of study for day for several students:

Study_hours= [2,3,6,5,4,7,8,2,3,4,5,3,2,2,4,5,6,8,9,1,2,9,5,6,8,7,8,9,10,10,9,8,5,3,2,1,20]

Now in the below box plot graph we can see the outlier. The outlier is a data that is different to the others, in this example the students have a maximum of 10 hours of study but one student studied 20 hours.

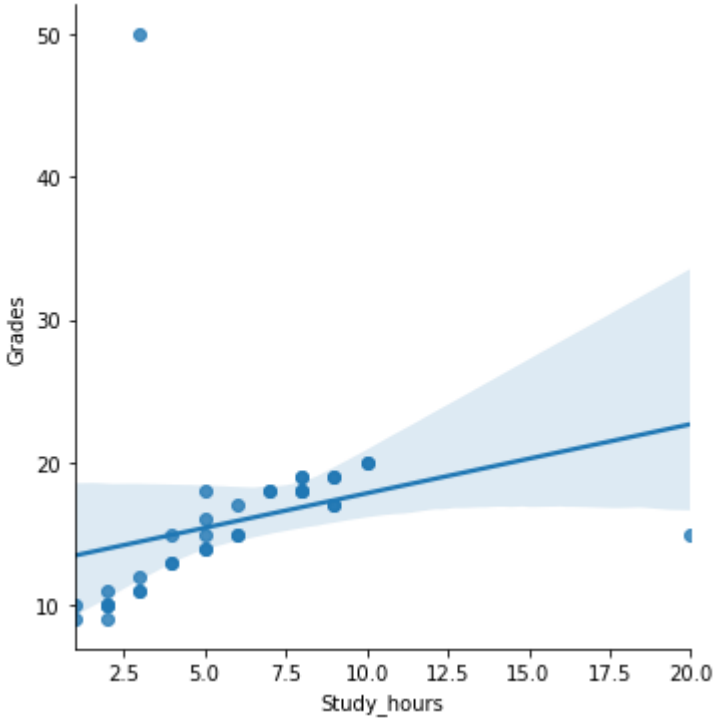
Graph: box plot, scatter plot



In this new example we have two variables, study hours and grades (1-20 scale):

study_hours= [2,3,6,5,4,7,8,2,3,4,5,3,2,2,4,5,6,8,9,1,2,9,5,6,8,7,8,9,10,10,9,8,5,3,2,1,20]
grades=[10,11,15,14,13,18,19,10,11,13,14,12,9,10,15,16,15,
18,19,10,11,19,15,17,18,18,19,17,20,20,17,18,18,50,10,9,15]

In the below scatter plot graph we can see the study hours outlier but also a new outlier related to the grades. The highest grade is 20 but one student has 50.



4. What is “Collaborative Filtering”?

Collaborative filtering systems stand out in multiple scenarios, especially when you want to establish a recommendation mechanism based on the similarity of users or content (products to be recommended). Collaborative filtering filters information, it uses the data collected from other users and interactions based on the idea that people who agreed in their evaluation of certain items are likely to agree again in the future.

There are two classes of Collaborative Filtering:

•**User-based,** When we talk about applying it to the search for similarity between users, we start from the basis that users who consume a service and have a similar experience could also do so with products that one of them has not yet tried. The technique consists of the construction of a matrix where each row represents a user and the columns represent the products that are subject to recommendation (the entire available catalog). The cells of the matrix determine the score that a user assigned to a specific product after its "consumption". The score is assigned depending on the parameters desired for the recommendation system, such as the rating that the user assigns to the product (star system, for example), number of times the user has visited or purchased the service, half session, etc. As these evaluations are recorded in the matrix, it is populated with information. When you have a sufficiently rich matrix, it is possible to apply algorithms that determine the similarity of the users with each other, correlating the rows of the matrix based on how similar the scores assigned in each of the columns are. As a result, we determine which users are similar to each other based on their scores. User-based collaborative filtering is not always the most appropriate because it presents a series of problems that should not be trivialized. The matrices that are built in this way are tremendously large to be handled by rows, since if the users are much more numerous than the available catalog of products, the correlation process will require a very high computational complexity and will require Big Data techniques to resolve.

•**Item-based,** this studies and use the similarity between the user rate items and the interaction with others item. With more users than items, each item tends to have more ratings than each user, so an item's average rating usually doesn't change quickly, this leads to more stable rating distributions in the model, so the model doesn't have to be rebuilt as often. When users consume and then rate an item, that item's similar items are

picked from the existing system model and added to the user's recommendations. Here it is possible to transpose the matrix to carry out content-based collaborative filtering, which consists precisely of the opposite, establishing similarity relationships between the products to order them in this way, reducing the computational requirements for its calculation, since the number of products is usually less than that of users. In item based explore the relationship between the pair of items (the user who bought Y, also bought Z). We find the missing rating with the help of the ratings given to the other items by the user.

Example of collaborative filtering are movies recommendation system and amazon recommendation system, in fact this technique is very used for several companies.

5. What is "Time Series Analysis"?

Time series analysis is a statistical technique that deals with time series data, or trend analysis, it accounts for the fact that data points taken over time may have an internal structure (trend and seasonal). Time series data is gathering observations obtained by constant measurements over time that data is well defined and it is in a series of particular time periods or intervals, also the time series must have sequential measurements and equal spacing between every two consecutive measurements an example are the sales for each month of the years or historic stock prices. The time series analysis is useful to obtain an understanding of the underlying forces and structure that produced the observed data and fit a model and proceed to forecasting, monitoring or even feedback and feed forward control.

There are types of times series:

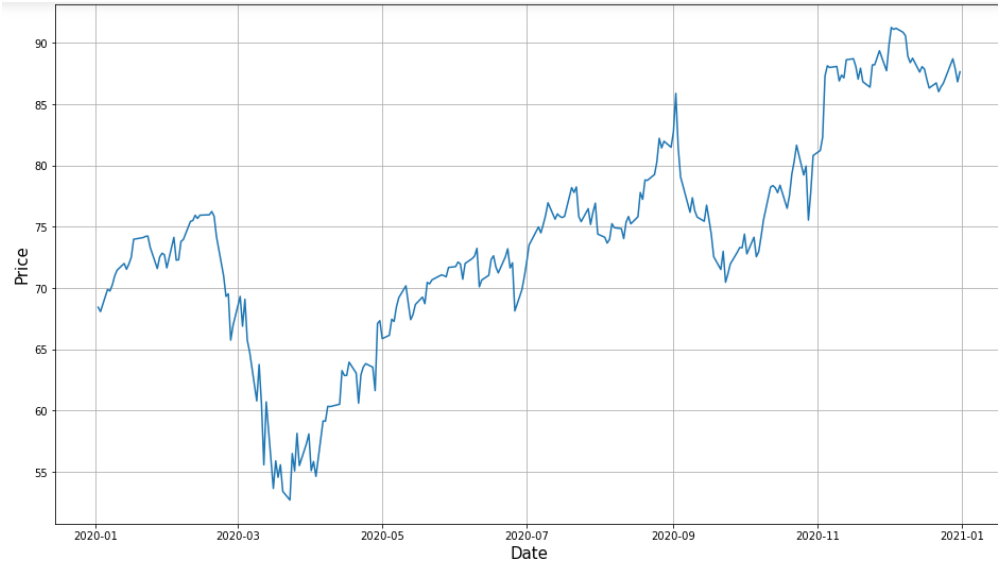
- Discrete or Continuous, based on the time interval considered for its measurement.
- Flow or Stock. In Economics, a data series is said to be of the flow type if it refers to a certain period of time (a day, a month, a year, etc.). For its part, a data series is said to be of the stock type if it refers to a specific date (for example, December 31 of each year). An example of flow type data would be the sales of a company since these will have a different value if the data is obtained after a week, a month or a year; For its part, the closing price of the shares of that same company would be a stock-type variable.
- Depending on the unit of measurement, we can find time series in euros or in various physical quantities (kilograms, liters, miles, etc.)
- Based on the periodicity of the data, we can distinguish time series of daily, weekly, monthly, quarterly, annual data, etc.

Time series have three components:

- ***Trend***, the trend of a time series is its long-term behavior or movement. It can be: uptrend, downtrend and horizontal.
- **Seasonality,** that are Periodic fluctuations of less than one year and recognizable every year, for example those related with the weather or the behavior of economic agents when the time of year varies.
- ***Cyclical patterns***, this component reflects recurring behaviors, although they do not have to be exactly periodic, with a period of more than one year. They usually show how the stages of economic boom follow one another with those of crisis, or at least slowdown.

There are innumerable applications that can be cited, in different areas of knowledge, such as business (sales forecasting, budgetary analysis), industry (process and quality control), economics (economic forecasting, stock market analysis), physics, geophysics, chemistry, electricity, demography(census analysis), marketing, telecommunications, transportation, etc.

The picture below is an example of time series graph:



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

Step 1: Understand the Business Issues. It is related to the definition of customer needs. This initial phase focuses on understanding the project objectives and also brief outline of the expectations is given to us with this we can identify the key objectives that the business is trying to uncover. This step is understanding what types of data are needed to answer the questions. This knowledge of the data is then converted into a definition of a data analysis problem and a preliminary plan designed to achieve the objectives. We have to examine the overall scope of the work, business objectives, information the stakeholders are seeking, the type of analysis and the deliverables (the outputs of the project). It's important to ask as many questions as we. Some questions to answer are: What decisions needs to be made? What information is needed to inform those decisions? What type of analysis can provide the information needed to inform those decisions?.

Step 2: Data understanding. Study and understanding of the data phase begins with the initial data collection, the data may collect internally through your CRM or from external sources, APIs, Google's public data, and third-party websites, and continues with activities to become familiar with the data, identify quality problems, discover preliminary knowledge about the data, and/or discover interesting subsets to form hypotheses about the hidden information. To do this there are a variety of tools to organize the data according its size like Excel, R, python, Alteryx, etc.

We should identify key variables to help categorize the data. When going through the data sets, look for errors in the data. These can be anything from omitted data, data that doesn't logically make sense, duplicate data, or even spelling errors. These missing variables need to be amended so you can properly clean your data.

Step 3: Prepare the Data (data preparation). This phase covers all the activities necessary to build the final data set (the data that will be used in the modeling tools) from the initial raw data. Once you have organized and identified all the variables in your dataset, it is possible to do task like selecting tables, records, and attributes, as well as data transformation and cleansing for the tools they model with activities like input and find missing, incomplete, or repetitive data or variables, create new broad categories to help categorize data that doesn't have a proper place, checking for outliers, imputing average data scores for categories where there are missing values, making sure that metrics like the mean, median, mode, and range make sense given the context. etc. Sometimes it is also necessary to convert the data into a format that is readable by data analysis tools.

Step 4: Perform Exploratory Analysis and Modeling. In this phase, the modeling techniques that are relevant to the problem are selected and applied and their parameters are calibrated to optimal values that's mean begin building models to test the data and find answers to the objectives given. There are typically several techniques for the same type of

data analysis problem. Some techniques have specific requirements on the form of the data. Some models are linear regressions, logistic regression, random forest model, decision trees, and boosted models.

Step 5. Validation. At this stage in the project, one or more models have been built that appear to be of sufficient quality from a data analysis perspective. Before proceeding to the final deployment of the model, it is important to thoroughly evaluate it and review the steps executed to create it, verify if we have the correct information for the deliverable, and compare the obtained model with the business objectives. A key objective is to determine if there are any important business issues that have not been sufficiently considered. At the end of this phase, a decision on the application of the results of the data analysis process should be obtained. Some question to answer are: the models work properly? Does the data need more cleaning? Did you find the outcome the client was looking to answer? If not, you may need to go over the previous steps again.

Step 6: Presentation and visualization of the findings. Generally, the creation of the model is not the end of the project when we have all your deliverables met, you can begin your data visualization. Depending on the requirements, the development phase can be as simple as generating a report or as complex as performing a periodic and perhaps automated data analysis process in the organization. Data visualization will be crucial in communicating the findings to the client, an interactive visualization tools are useful in illustrating the conclusions to clients. Being able to tell a story with your data is essential. Telling a story will help explain to the client the value of your findings.

7. What are the characteristics of a good data model?

Data modeling is understood as the process of designing and creating data models intended to serve in a given information system. A complex and variable process depending on organizational needs and established corporate objectives.

•Allows to get the objectives and expected results.
•Throws away everything except the necessary to understand the underlying pattern that it is going to be reveal.
•Measure its success on at least two axes: predictive and explanatory power.
•It doesn’t tend to be overfitted or underfitted.
•A good data model is intuitive to develop upon.
•A great data model is one that can evolve and support new business cases.
• Data in a good model can be easily consumed.
• Large data changes in a good model are scalable.
• A good model provides predictable performance.
• A good model can adapt to changes in requirements.
• Facilitates speed and efficiency in the design and creation of databases
•Creates consistency in data documentation and system design throughout the organization
• It makes the display of information clean and easy to understand.
• Organized data can be subjected to quality control to avoid errors.
• Allows the structuring and organization of data to be arranged and distributed in a highly precise order, which favors its handling and processing.
• Helps identify duplicate and unnecessary data.
• Indicates if certain data is missing from the modeling.
• Ensures that a company's system design is consistent.
• Makes sense of the voluminous and messy data that can come from various sources in an organization.
• Optimizes communication between developers and business intelligence systems.
• Allows the design of high-quality databases with the aim of helping to better create applications

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

Univariate analysis, this analysis works with univariate data, this data type consists of a single variable. When the analysis presents characteristic by characteristic, in isolation, we will be in the presence of a univariate statistical analysis. It is possible found two kinds of variables categorical and numerical. Univariate data analysis is the basic and primary form of analysis because the information deals with a single data or characteristic regardless of any other characteristic. The purpose of univariate analysis is to understand the distribution of values for a single variable it doesn't search for causes or relationships.

The patterns can be identified using measures of central tendency (mean, median and mode), dispersion (range, variance and standard deviation), quartiles (interquartile range), data extension (range, minimum, maximum). This data can be described through: frequency distribution tables, bar charts, histograms, frequency polygons, and pie charts.

Example: An example can be a survey about teens and technology, an univariate analysis only works with one variable that can be the hours of children play videogames, the type of electronic electronics devices they use (PSP, Wii, computer, etc), the city, the age, the hours of study, hours in social media, weight, dietary habits, problems with family, time spend with their friends etc. We can obtain the average hours of play and social media, in which country live most but we don't going to find a relationship or explain a causality.

Demonstrative Example:
We have a dataset about average salary from Peru by year but this is a univariate analysis so I am going to use the data from 2020 to describe it but not to find relations or causality:

```
import pandas as pd
data=pd.read_excel('peruvian salaries.xlsx')
sal_2020=data[[2020]]
sal_2020.head()
```

So we have the variable sal_20 with following head:

	2020
0	992.895413
1	1057.212206
2	1004.480868
3	1530.315818
4	1095.425114

We can describe the data to found the maximum and minimum value, the quartiles, the mea n, the standard deviation. So we found that the average salary in Peru is 1144 PEN.

```
sal_2020.describe()
```

	2020
count	27.000000
mean	1144.310802
std	259.448991
min	669.030011
25%	987.915503
50%	1118.085071
75%	1264.370499
max	1711.051929

We can do also a frequency distribution table with intervals of salaries:

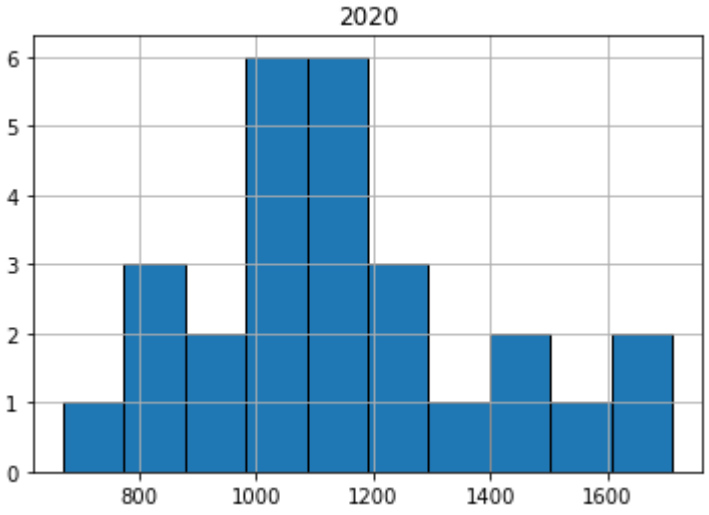
```
bins = [600, 900, 1200, 1500, 1800]
sal=sal_2020[2020].to_list()
categories = pd.cut(sal, bins)
categories
pd.value_counts(categories)
```

```
(900, 1200]     13
(1200, 1500]     6
(600, 900]       5
(1500, 1800]     3
dtype: int64
```

Also we can graph the data:

With histogram to understand the distribution that is asymmetric and right skewed:

```
sal_2020.hist(edgecolor='black', linewidth=3)
plt.show()
```



Bivariate analysis, Bivariate data come from the simultaneous observation of two variables (X, Y) in a sample of n individuals. The data will be pairs of values, numeric or non-numeric, used to describe the two variables together or a variable in function of the other. In studies of relationships between variables, one of the two variables plays a more important role than the other, this will be the dependent variable that we will denote by y, whose behavior we will try to describe in terms of another variable x that we will call independent or explanatory variable. The data can be shown in a double entry table.

The bivariate analysis has the following types:

- Bivariate Analysis of two Numerical Variables (Numerical-Numerical): Scatter Plot represents individual pieces of data using dots. These plots make it easier to see if two variables are related to each other. The resulting pattern indicates the type (linear or non-linear) and strength of the relationship between two variables. Linear Correlation, this represents the strength of a linear relationship between two numerical variables. If there is no correlation between the two variables, there is no tendency to change along with the values of the second quantity. Here we verify the value of r (correlation coefficient) is the specific measure that quantifies the strength of the linear relationship between two variables in a correlation analysis and r-squared (Coefficient of determination) the statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. The value of the coefficients is always between -1 and 1 where -1 denotes perfect negative linear correlation and +1 denotes perfect positive linear correlation and zero denotes no linear correlation. Some statistical tools are Pearson correlation.

- Bivariate Analysis of two categorical Variables (Categorical-Categorical): the Chi-square Test is used for determining the association between categorical variables. It is calculated based on the difference between expected frequencies and the observed frequencies in one or more categories of the frequency table. A probability of zero indicates a complete dependency between two categorical variables and a probability of one indicates that two categorical variables are completely independent. Other statistical tools are Fisher test, McNemar test, V cramer coeficient and Binomial test.

- Bivariate Analysis of one numerical and one categorical variable (Numerical-Categorical): Z-test (If the probability of Z is small, the difference between the two averages is more significant) and t-test (the measure must be less than 0.05) calculate if the difference between a sample and population is substantial. Some statistical tools to use are T-student, U-Mann Whitney, Kruskal-Wallis, Friedman test and the ANOVA test (Analysis of variance).

Example: An example can be try to find correlation between variables of teens and technology survey like hours in computer and hours of study, time they spend with friend with hours in electronic devices, time they spend playing videogames and food habits, time they spend playing videogames and family problems. Determinate the relation between the type of electronic devices and sex. Also find and compare averages hours of playing videogames by city, age and sex.

Demonstrative example:

Now we are going to work with two variables 2020 and geographic zone, for this bivariate analysis I am going to analyze the mean difference between geographic regions and its average salaries:

```
geo_salry=data.iloc[:,[0,14]]
geo_salry.head()
```

	Ámbito geográfico	2020
0	Amazonas	992.895413
1	Áncash	1057.212206
2	Apurímac	1004.480868
3	Arequipa	1530.315818
4	Ayacucho	1095.425114

How are many zones we are going to select four zones with the best salary and worst. It is just an example to show the difference between one and two variables because in the theory we look the diverse analysis that for time reasons I didn't do.

```
maxi=geo_salry.nlargest(n=2, columns=2020)
mini=geo_salry.nsmallest(n=2, columns=2020)
frames=[maxi, mini]
regions=pd.concat(frames)
```

	Ámbito geográfico	2020
14	Lima Metropolitana 1/	1711.051929
18	Moquegua	1693.717084
8	Huancavelica	669.030011
21	Puno	809.783383

The zone with the best salary is Lima, the average salary is 1711.05, and Huancavelica has the worst salary. The salary isn't near from the basic salary that is 1000 PEN.

Multivariate analysis, Multivariate analysis is required when more than two variables have to be analyzed simultaneously. Multivariate analysis brings together statistical methods that focus on simultaneously observing and processing different statistical variables to obtain relevant information. The reason to use it lies in a better understanding of the phenomenon under study, obtaining information that univariate and bivariate statistical methods are unable to obtain.

Some methods of multivariate analysis are: Additive Tree, Canonical Correlation Analysis, Cluster Analysis, Correspondence Analysis / Multiple Correspondence Analysis, Factor Analysis, MANOVA, Multidimensional Scaling, Multiple Regression Analysis, Partial Least Square Regression, Principal Component Analysis / Regression / PARAFAC, Redundancy Analysis.

Example: determinate if some teenagers are going to develop addiction or not. Realize a cluster to classify by zones to develop health politics for teenagers, or in which zone I am going to focus to advertise my new videogame. Determinate which variables are related with the quantity of playing videogames hours and being in social media.

9. What is a Linear Regression?

Linear regression is a field of study that emphasizes the statistical relationship between two continuous variables known as the predictor and response variables, when there is more than one predictor variable, it is converted to multiple linear regression but there are also other types of linear regression: logistic regression, ordinal regression, multinomial

regression, discriminant analysis. It adapts to a wide variety of situations. In social research it is used to predict a wide range of phenomena such as economic measures and aspects of human behavior. In market research to determine which media to invest in, predict the number of sales of a product.

Simple and multiple regression are used to explore and quantify the relationship between a dependent variable or response (Y) and one or more dependent or predictor variables (X_1, X_2...X_k), as well as develop a linear equation for predictive purposes.

Simple linear regression consists of generating a regression model (equation of a straight line) that allows explaining the linear relationship that exists between two variables.

$$Y = \alpha + \beta X$$

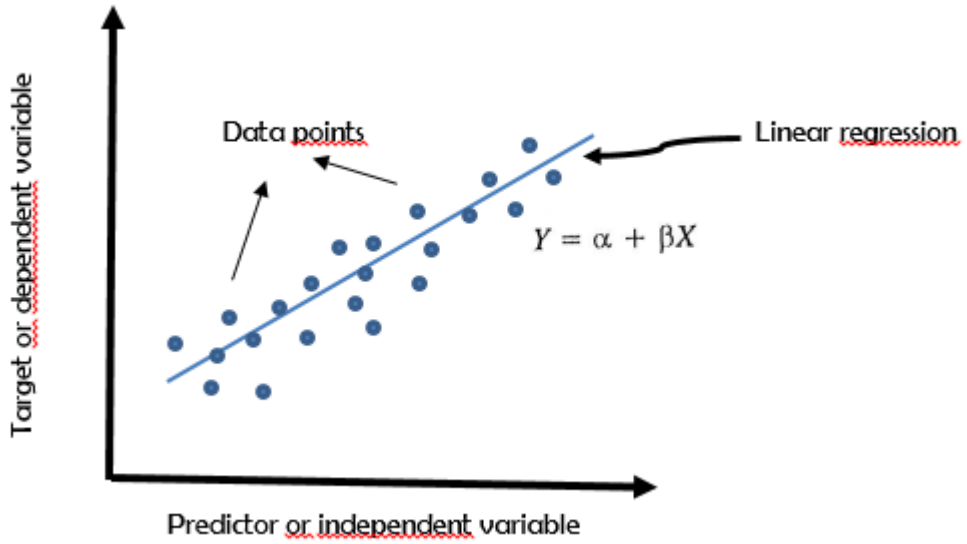
Y is the dependent or response variable;
X represents the explanatory, independent or response variable;
α is the ordinate at the origin (intercept) and β the slope are the parameters of the model, which measure the influence that the explanatory variables exert on the regression.

The multiple linear regression equation is:

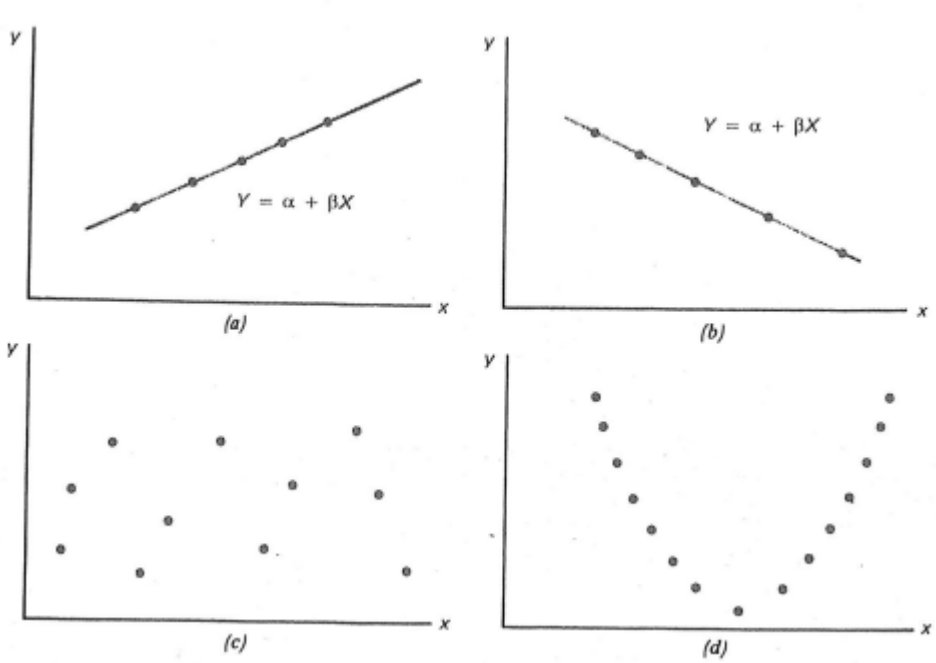
$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

y = the predicted value of the dependent variable
β_0= the y-intercept (value of y when all other parameters are set to 0)
β_1x_1= the regression coefficient (β_1) of the first independent variable (X_1) (the effect that increasing the value of the independent variable has on the predicted y value)

The linear regression is represented by the scatter plot graph:



Analyzing the scatter plot and regression line we can observe the following:



a)Perfect positive correlation: p=1, β>0, b) perfect negative correlation p=-1, β<0, c) no correlation, p=, d) there is a relation but it isn't lineal p=0.

To work with linear regression is important to understand the linear correlation, there are three basic types of coefficients: Pearson coefficient, Spearman Rank (rho) and Kendall rank (tau). The value of r (correlation coefficient) is the specific measure that quantifies the strength of the linear relationship between two variables in a correlation analysis and r-squared (Coefficient of determination) the statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. All of them vary between +1 and -1. Being +1 a perfect positive correlation and -1 a perfect negative correlation. They are used as a measure of association strength (effect size):

0: null association.

0.1: small association.

0.3: medium association.

0.5: moderate association.

0.7: high association.

0.9: very high association.

In addition to the value obtained for the correlation coefficient, it is necessary to calculate its significance. Only if the p-value is significant can it be accepted that there is a correlation, and this will be of the magnitude indicated by the coefficient. No matter how close the value of the correlation coefficient is to +1 or −1, if it is not significant, it must be

interpreted that the correlation of both variables is 0, since the observed value may be due to simple randomness. The p-value is less than 0.05.

Example:

We are going to find the realtion between working hours and a rate of satisfaction in work:

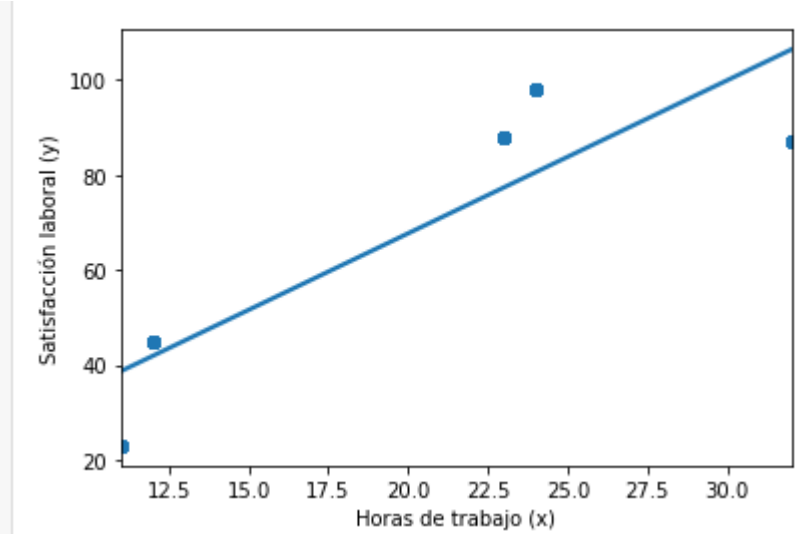
The correlation to determinate is pearson correlation.

```
import numpy as np

import scipy.stats



datacor=pd.read_excel('ejemploreg.xlsx')

x = datacor['Horas de trabajo (x)']

y = datacor['Satisfacción laboral (y)']



scipy.stats.pearsonr(x, y)
```

```
(0.8646553191071562, 6.769665224311492e-20)
```

The correlation is strong because the values is 0.86, according the explained above. Also the p-value is 0.000 this value is significative. So there is relation between the hours in work and work satisfaction.

We can do the scatter plot with linear regression:



Then we can do a predictive equation to predict values.

10. In terms of modelling data, what do we mean by Over-fitting and Under-fitting?

Overfitting, is a modeling error that occurs when a model is excessively complex, such as having too many parameters relative to the number of observations or data points when a model is trained with much data, it begins to learn from noise and inaccurate data inputs in our dataset. A model that has been overfitted has poor predictive performance because it overreacts to minor fluctuations in the training data. Models that incur in overfitting tend to have high variance. When there is overfitting we cannot model the training data or generalize to new data. So the model does not categorize the data correctly, due to too much detail and noise. An example when we develop a machine learning model, we try to teach it how to achieve a goal: detect an object in an image, classify a text based on its content, speech recognition, etc. To do this, we start from a database that will be used to train the model, that is, to learn how to use it to achieve the desired objective. However, if we don't do things right, the model may consider as validated only the data that was used to train the model, without recognizing any other data that is slightly different from the initial database because ovefitting means that the model loses the ability to generalize since instead of capturing the general patterns underlying the training set, it pays too much attention to all the details, including the superfluous and inconsequential ones that do not affect in any way on the training set. the prediction. In this way, what the model does is memorize, not learn.

Some ways to solve and avoid overfitting are:

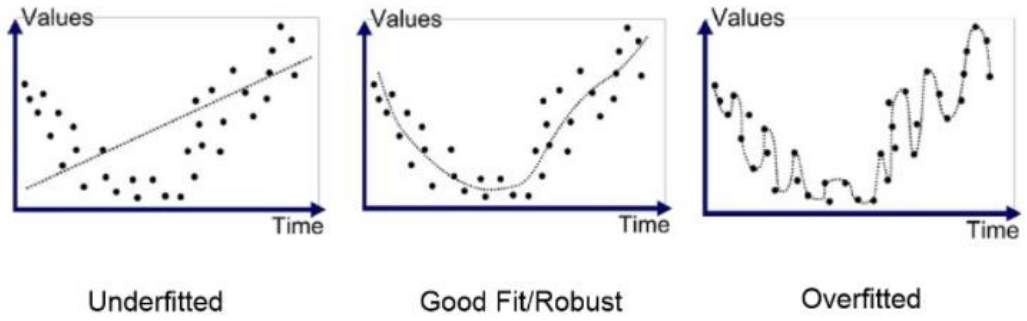
- Training with more data, this technique may not work every time. Basically, it helps the model to better identify the signal. But in some cases, increasing data can also mean feeding more noise into the model. When we train the model with more data, we need to make sure that the data is clean and free of randomness and inconsistencies.
- Use simpler models: One of the main reasons why overfitting occurs is because we use models that are too complex for our data.
- Early stop, when the model is being trained, you can measure the performance of the model based on each iteration. We can do this up to a point where iterations improve the performance of the model. After this, the model overfits the training data as the generalization weakens after each iteration. Early stopping refers to stopping the training process before the learner passes that point.
- Elimination of functions, although some algorithms have an automatic selection of functions. For a significant number of those that do not have built-in feature selection, we can manually remove some irrelevant features from the input features to improve generalizability.
- Assembly, This technique basically combines predictions from different Machine Learning models. Two of the most common methods for assembling are: Bagging attempts to reduce the possibility of overfitting models, Drive attempts to improve the predictive flexibility of simpler models.
- Cross-validation is a powerful preventive measure against overfitting.The idea is use your initial training data to generate multiple mini-train test splits. Use these divisions to adjust your model.Cross-validation allows you to fit hyperparameters only against their original training set. This allows you to keep your test set as a truly invisible data set to select your final model.
- Regularization: It is a way of penalizing complexity in models, typically at the parameter level.
- Remove irrelevant “features”.

Underfitting, is a modelling error that occurs when a statistical model or machine learning algorithm fails to capture the underlying trend in the data when the model or algorithm does not fit the data well enough, the model is very simplistic, insufficient to capture the nuances, particularities and complexities in the data. Intuitively, this occurs specifically if the model or algorithm exhibits low variance but high bias. The models that are victims of underfitting have high bias, which translates into an almost obstinate robustness to variations in the data, to

the point of not being able to detect the characteristic nuances of the data on which they work. This would occur, for example, when fitting a linear model to nonlinear data. Such a model would also have poor predictive performance.

To detect if our model is underfitting we use the same methods of the overfitting.

Some ways to solve and avoid the underfitting are:
• Use more complex models.
• Collect more data.
• Synthesize more data (for example, through data augmentation).
• Use cross-validation to make better use of available data.
• Reduce regularization, if it is being used.
• Eliminate irrelevant “features” to mitigate “the curse of dimensionality”, which establishes that, approximately, for each “feature” in our data, we will need 10 times more training instances to be able to fit a good model.



Source: https://julienharbulot.com/overfitting.html

To detect the both:
There are several ways to detect if our model is overfitting the data. The best method is to take a part of the training data to validate the model's performance, acting as new data.
We will then measure the performance of the network on both data sets, using the relevant metric (for example, accuracy), and compare them. If the gap between the performance on the training data and the test data is very wide, it means that our model is effectively “overfitting”, that is, memorizing, not learning.In these cases it is more useful and intuitive to plot the values of the target metric, as well as the loss function, over the epochs/iterations. Some methods are:

- The bias-variance trade-off because the variance represents how much the function changes when changing the training set. If it changes a lot, we say that the statistical model has a high variance, so it surely suffers from overfitting since it is capable of perfectly modeling the training data, but when it generalizes to data that it has never seen, then it fails. The bias can be defined as the opposite. If, when using different training sets, the function remains practically the same, then the model has a low variance and a high bias.
- Learning curves is one of the best methods to diagnose our model of possible problems of overfitting (high variance) or underfitting (high bias). In a typical graph of learning curves we have an error metric on the ordinate axis, for example, the MSE (Mean-Squared Error) and different sizes of the training set on the coordinate axis.The learning functions will indicate how the model error varies with the size of the learning data set. In the case of overfitting or high variance, the graph will show how there is a large gap between the validation data and the training data. This is because the model fits the training data very well, so the error in the training set will be very low. However, he is unable to generalize. For this reason, the error in the validation set will be much larger. In the following graph we can see the typical learning curves of an overtrained model. When we have a high bias or underfitting, the gap between the two functions is very small. Also, the error of the validation set and the training set is high. This indicates that it is a very simple model that does not fit the data well. In this case it

would be necessary to add more training data or increase the model training time.