Date: 22/09/2022

Final Report about the Summer Training Conducted

By
Mostafa Mohamed Ismail Ahmed
ID
20191613931
Program
Computing and data science

Introduction:

The main goal for this training course is:

- Strengthen graduates' ability to work in the local market by enhancing their capabilities and skills.
- Create and select features for predictive modeling
- Deploy and manage analytical models under production.
- Evaluate and select the best model based on business needs

ACKNOWLEDGEMENT

- The training received financial and technical support from Sprint.

Training Name:

- Data Science Summer Program

The topics of the course:

- Agile & Scrum
- SW Configuration Management
- Data Science
  - Introduction to Data Analytics
  - Data Science Toolbox
  - Data visualization
  - Mathematical Basics
  - Data Modeling
  - Time Series Analysis
  - Data Acquisition & Understanding
  - Advanced Statistical Models
  - Capstone Project
  - Recursive Partitioning

Date and Period:

Training for the groups:

- The training course started on 30/7/2022 and completed 10/9/2022.
- The duration of the training is 60 hours for 6 Weeks.

Place of training:

- The activity took place Online by Sprints.

The instructors:

- 3 trainers participated in the activity as follows:

    1. Dr. Kareem Abdallah taught 60 hours

Total 60 hours

The target groups:

50 representatives from Alexandria University participated in this activity.

Methodology of training:

The training focused on:

- Self-Learning Through Interactive and Creative Learning.
- with the added convenience of a learning experience tailored to my schedule. With courses available from Sprints platform, and flexible timetables to suit almost every lifestyle.
- traditional classroom teaching model.
- Assignments And Tasks method
- Presenting the subjects then discuss it on the lecture and let the trainees participate.

Subject of the course:

- **Agile & Scrum:**
    - Why Agile?
    - Agile Foundation
    - Agile Deep Dive
    - Scrum Foundation
    - Scrum Deep Dive
    - Agile Challenges
    - Agile vs PM

- **SW Configuration Management:**
    - Software Configuration Management
    - Git & GitHub

- **Data Science**
    - **Introduction to Data Science:**
        - Data Science: Why all the Excitement?
        - How it all Began?! An Intro to Big Data
        - Data Analytics Life Cycle
        - Different Data Disciplines
        - AI, Machine Learning & Deep Learning
        - Machine Learning Overview
        - Data Science as a Career!
    - **Data Science Toolbox:**
        - Introduction to Python
        - Practical Examples

- **Mathematical Basics:**
  - Introduction to Statistical Analysis
  - Introduction to Inferential Statistics
  - Central Limit Theorem
  - Confidence Interval
  - Probability Principles - Intro
  - Probability Principles - Random Variables
  - Probability Principles - Probability Functions
- **Data Modeling - Time Series Analysis**
  - Introduction to Time Series Analysis
  - Plot Analysis
  - Modelling with Autoregressive Models
  - Stationarity
  - MA
  - ARIMA
  - SARIMA
  - Decomposition
  - Facebook Prophet
  - Other Time Series Modelling Algorithms
  - Other Forecasting Algorithm

The training details

- **10 Hours per week:**
  - **3 Hours on live sessions**
  - **3 Hours on recorded videos**
  - **4 Hours on Tasks and studying**
- **WEEK 1:**
- Live discussion on several career paths in data science. Become acquainted with one another and form project teams. describing the course of events and the road map
- On the platform, there were two crash courses.
  - **Agile & Scrum:**
    - Why Agile?
    - Agile Foundation
      - Agile Manifesto 1
      - Agile Manifesto 2
      - Agile Principles 1
      - Agile Principles 2
    - Agile Deep Dive
      - Self-Organized Teams 1
      - Self-Organized Teams 2
      - Time Boxing
      - Agile Practices
    - Scrum Foundation
      - Scrum
      - Scrum Lifecycle
      - Scrum Documents
      - Relative Estimation
    - Scrum Deep Dive
      - Sprint Planning
      - Task Board & Burndown chart
      - Daily Scrum

        - Sprint Review Meeting
        - Sprint Retrospective Meeting
    - Agile Challenges
    - Agile vs PM
- **SW Configuration Management:**
    - Software Configuration Management
        - Introduction
        - Outlines
        - Software Development Life Cycle (SDLC)
        - What is Configuration Management
        - SW Configuration Management Activities
        - SW Configuration Management Tools
        - What is Version Control
    - Git & GitHub
        - Introduction to Git and GitHub
        - Working copy, Staging and Commit
        - Branching and Merge
        - Git workflow
        - Git Basic commands
        - Git in action
        - Branch & Merge in Action
        - Merge Conflicts
        - GitHub in Action
        - Remote Repo workflow
        - More into GitHub
        - Git GUI Clients
        - UI Tools in Action

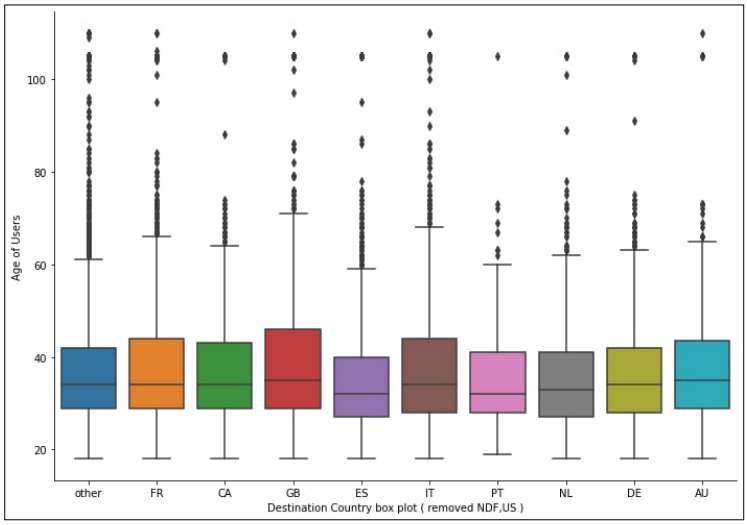
➢ **WEEK 2:**

❖ Recorded videos:
- Introduction to Data Analytics
    - Data Science: Why all the Excitement?
    - How it all Began?! An Intro to Big Data
    - Data Analytics Life Cycle
    - Different Data Disciplines
    - AI, Machine Learning & Deep Learning
    - Machine Learning Overview
    - Data Science as a Career!

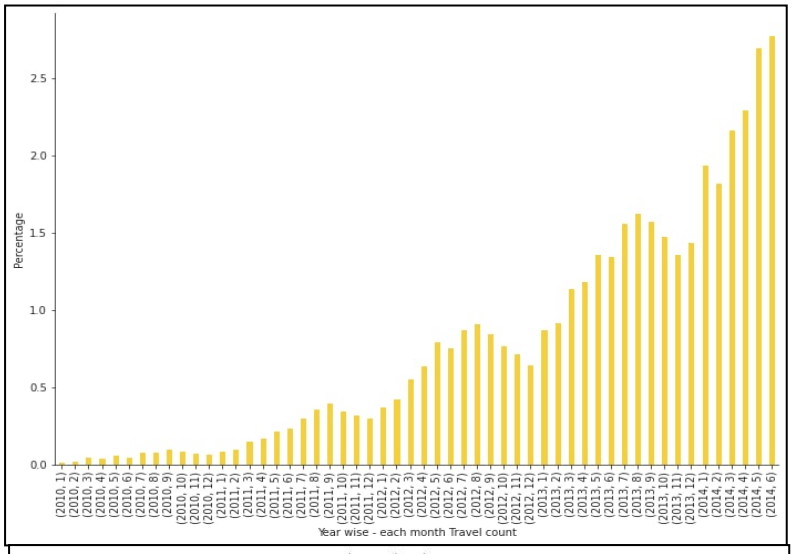
❖ Live session:
it's crucial to get discover the hidden insights within the data at hand: distributions, discovering data quality problems, detecting outliers, finding correlations, discarding redundancies, getting subsets of interest from main data sources
- Intro to EDA & Visualization
- Summary Statistics
    - in statistics, exploratory data analysis in an approach to analyzing data sets to summarize their main characteristics, often with visual methods.
- Single Variable Visualization
    - Histograms
    - Box Plots
- Multivariate Visualization
- Most Popular Visualization Tools

❖ Task:

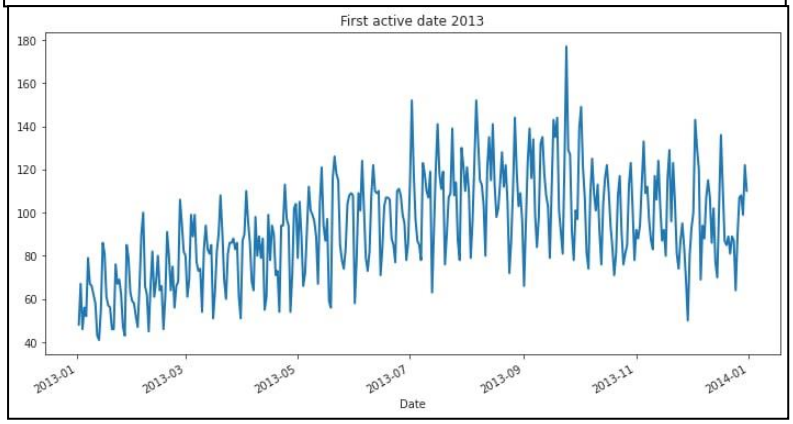
- **NotebookURL:**https://colab.research.google.com/drive/1vCy4oTcviq2JHZkfXiMYVRFLvheNNo-R?usp=sharing
  - Airbnb data set on Kaggle to visualize some aspects and get insights from it.



Users booking for countries Spain, Portugal and Netherlands tend to be younger where as Users booking for Great Britain tend to be older



We can see from the graph that each year has similar trend that in 7,8,9 months the chance of booking is higher than the rest of the year.



If we see month wise activity of the users then the peak months were July, August and October. On the other hand, least active month was December

- **WEEK 3:**
- Recorded videos:
  - Data Science Toolbox:
    Python for Data Analysis is concerned with the nuts and bolts of manipulating, processing, cleaning, and crunching data in Python. It is also a practical, modern introduction to scientific computing in Python, tailored for data-intensive applications.

        - Introduction to Python
        - Practical Examples
    - Mathematical Basics:
        - Introduction to Statistical Analysis
        - Introduction to Inferential Statistics
        - Central Limit Theorem
        - Confidence Interval
        - Probability Principles - Intro
        - Probability Principles - Random Variables
        - Probability Principles - Probability Functions
- Live session:
    - Intro to Big Data
        - The Early Days
        - Distributed Systems
        - Evolution
        - Common Use Cases - Unlocked
        - Challenges with Distributed Systems
        - Hadoop Ecosystem
        - HDFS Cluster Operations

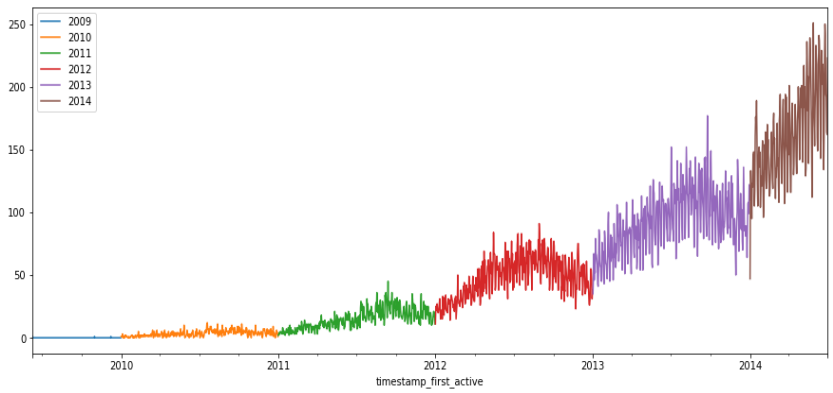
- **WEEK 4:**
- Recorded videos:
    - Data Modeling - Time Series Analysis
        - Introduction to Time Series Analysis
            - The main target is to determine a model that describes the pattern of the time series.
            - Univariate Time Series
                - A univariate time series is a sequence of measurements of the same variable collected over time. Typically, the measurements are made at regular time intervals.
        - Plot Analysis
        - Modelling
            - AR Models
                - This is called an AR (1) model, standing for autoregressive model of order 1. The order of the model indicates how many previous times we use to predict the present time.
            - Stationarity
                - For a series to be modelled, it must exhibit a "stationarity" property. An interesting property of a stationary series is that theoretically it has the same structure forwards as it does backward. To create a (possibly) stationary series, one solution is to examine the first difference yt= xt-xt-1 This is a common method for creating a de-trended series and thus potentially a stationary series.
            - MA
                - moving average model is based on the past error (multiplied by a coefficient), added to the mean within the MA window.
            - ARIMA
                - ARIMA models are models that may possibly include autoregressive terms, moving average terms, and differencing operations.
            - SARIMA

- (Seasonal Autoregressive Integrated Moving Average)Seasonality in a time series is a regular pattern of changes that repeats over S time periods.

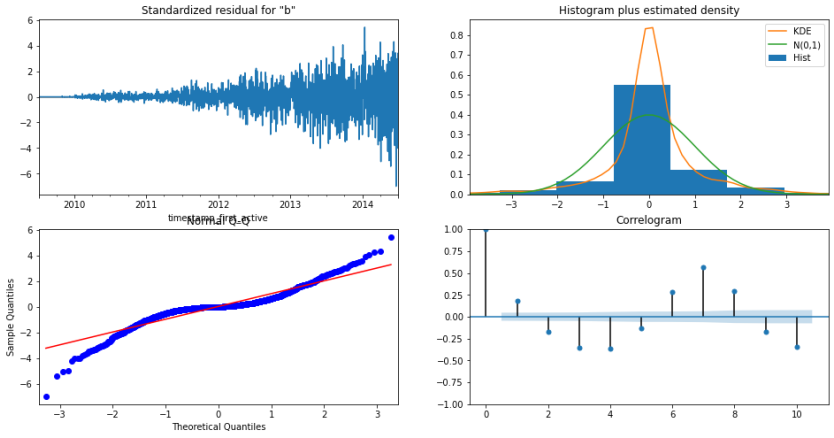
❖ Live session:
- Essential Terminologies
  - Comparing Different Analytics Types
    - Descriptive Analytics, Diagnostic Analytics, Predictive Analytics and Prescriptive Analytics
  - Business Intelligence Vs Data Science
    - Data Science, as used in business, is intrinsically data-driven, where many interdisciplinary sciences are applied together to extract meaning and insights from available business data, which is typically large and complex.
    - Business Intelligence (BI) helps monitor the current state of business data to understand the historical performance of a business.
  - Data Mining
    - data mining is mainly about finding useful information in a dataset and utilizing that information to uncover hidden patterns.
  - Structured vs Unstructured Data
    - Structured data is highly organized and formatted in a way so it's easily searchable in relational databases.
    - Unstructured data has no pre-defined format or organization, making it much more difficult to collect, process, and analyze.

❖ Task:
- about time series based on Airbnb data set from Kaggle to determine which country has the highest travelers over 5 years from 2009 to 2014 using ARIMA and ARMA.

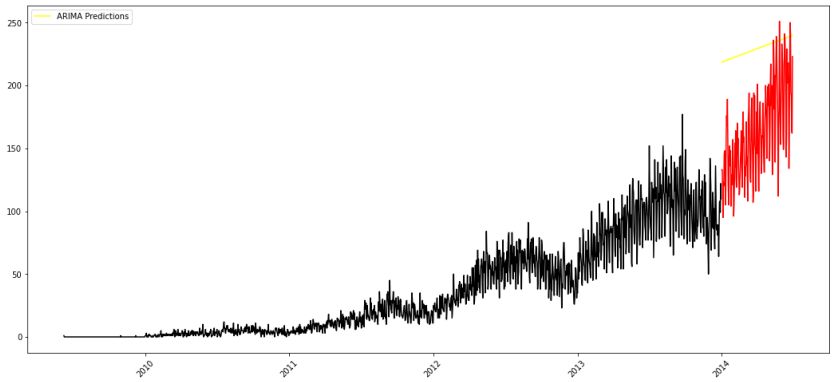
❖ **NotebookURL:**https://colab.research.google.com/drive/186mtndKzNyIizaMnApEttKGryCbvtQp_?usp=sharing



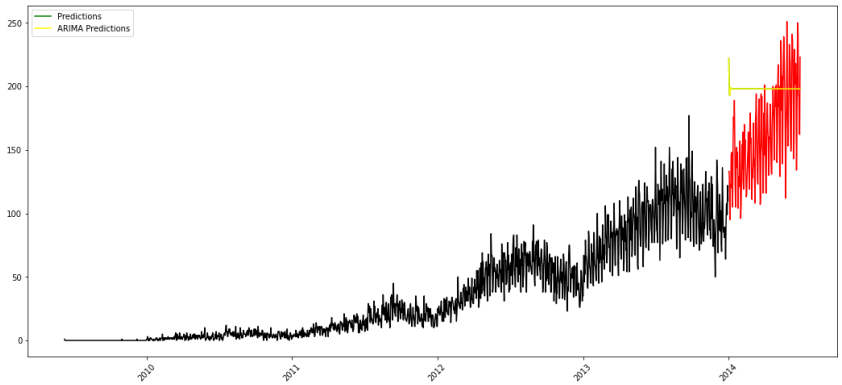
booking are always low at the beginning of the year and high at the end of the year. There is always an upward trend within any single year with a couple of low months in the mid of the year.



It is not perfect, however, our model diagnostics suggests that the model residuals are near normally distributed.



Arima Model with RMSE of 74.52



Arima Model with RMSE of 50.97

➢ **WEEK 5:**

❖ Recorded videos:

- Data Modeling - Time Series Analysis
  - Decomposition
    - Time series decomposition involves thinking of a series as a combination of level, trend, seasonality, and noise components
  - Facebook Prophet
    - The model is based on decomposing the time series into three main model components: trend, seasonality, and holidays. There are two implemented trend models:

      1)A saturating growth model. 2)Linear Trend with Changepoints
  - Other Time Series Modelling Algorithms
  - Other Forecasting Algorithm

- Amazon Gluon Time Series (GluonTS), an Open-Source Time Series Modeling Toolkit. Deep Learning Based Algorithms: Especially Recurrent Neural Networks (RNN) Algorithms, like LSTM & GRUs

❖ Live session:
- Machine Learning
  - Intro
  - Supervised ML
    The task of inferring a function from labeled data.
    - Classification
      - Linear Classifiers
      - Non-Linear Classifiers
    - Regression
  - Unsupervised ML
    Find hidden structure in unlabeled data

❖ Task:
- Task to make a prediction model on Airbnb data set using decision tree and KNN
- NotebookURL:https://colab.research.google.com/drive/1K91jB_2oxJbSxWtHBl73QiBdiwZ0LFO7?usp=sharing

apply Decision Tree algorithm

```
[ ] from sklearn.tree import DecisionTreeClassifier
    from sklearn.metrics import accuracy_score
    model = DecisionTreeClassifier(criterion='entropy',max_depth =9 , random_state=0)
    model.fit(x_train_sc, y_train)
```

apply KNN

```
[ ] from sklearn.neighbors import KNeighborsClassifier
    knn_model = KNeighborsClassifier(11)
    knn_model.fit(x_train_sc, y_train)
    y_pred= knn_model.predict(x_train_sc)
    print(accuracy_score(y_train,y_pred)*100)

    70.3169463251889
```

```
[ ] from sklearn.neighbors import KNeighborsClassifier
    knn_model = KNeighborsClassifier(11)
    knn_model.fit(x_train_sc, y_train)
    y_pred= knn_model.predict(x_test_sc)
    print(accuracy_score(y_test,y_pred)*100)

    69.67106769442113
```

➢ **WEEK 6:**

❖ Live session:
- Machine Learning
  - Feature Engineering
    ➢ It enables the machine learning algorithm to train faster
    ➢ It reduces the complexity of a model and makes it easier to interpret.
    ➢ Example: Sequential Forward Selection.
    - Transformation
    - Dimensionality Reduction
      - Example: PCA
    - Feature Selection

    - Example: Correlation Methods, Wrapper Methods
  - The next portion of the session was spent talking about our submissions for the assignments, assessing our methods for using the algorithms, and discussing how to get more knowledge and improve accuracy.

Positive aspect of the training:

- The participants were enthralled by the subject.
- The instructors utilized a democratic system to run their classes.
- The participants have accumulated a reasonable level of skill and experience.
- All the participants' relationships were harmonious.
- All sessions were held on time and all materials were available.

Method of training:

Using the following methods:

- Interactions between individuals or groups.
- groupings for discussion.
- Exercises for individuals and groups based on real-world experience.

Recommendations:

- in future training, participants hope to receive additional training in Data Science, with a particular focus on youth sectors.

Evaluation:

- Following the training, participants were asked to complete evaluation forms to gauge its efficacy. And I'm happy with the result.