Project 1: Data Analytics & Visualization on MS Excel

Project : **Black Friday Sales Analysis to understand Consumer Behavior**

Difficulty Level: Intermediate

Tools to Use: Microsoft Excel 2019 or newer

Aim: Analyzing the Black Friday sales to understand the customer behavior based on their demographics and product details.

Objectives:

1. Create a Buyers Demographics dashboard with charts showing the count of customers in different categories in each demographic's variable.
2. Create a Spending Dashboard to showcase the amount of money spent by different demography of buyers. This can include charts such as marital status and age groups vs total spending, occupation and gender vs total spending, etc.
3. The visuals should include different types of charts such as pie chart, clustered column chart, stacked column chart, table etc.
4. On a Conclusion dashboard mentioning your observation and findings from your analysis. For example, top buyers' profile, top products, top drivers behind customer purchasing decisions, factors that drive a particular female consumer segment to buy, cities which are likely to spend most, etc.

Dataset: Black Friday Sales

Skills to Use (Mandatory):

1. Exploring data in workbook and worksheets
2. Creating tables to enable Table Design tab
3. Using conditional statement for labelling (0 = Unmarried, 1 = Married)
4. Summarizing data with Pivot Tables
5. Using Pivot Chart Analysis
6. Adding Slicers and Filters to charts
7. Creating dashboards with suitable titles

Skills to Use (Optional):

1. Dashboard designing choosing a specific design template
2. Adding a company specific image/logo in the dashboards
3. Creating an Index dashboard with links to all other dashboards
4. Adding Previous and Next buttons in each dashboard

Resource Links:

1. Data Analysis on Excel - LinkedIn Learning (Basic data visualization and statistical analysis on Excel) : https://www.linkedin.com/learning/learning-excel-data-analysis/
2. Exploring data in workbook and worksheets
   Article: Difference between workbook and worksheet
3. Creating tables to enable Table Design tab
   Video: https://www.youtube.com/watch?v=Du73CPqWGQw
4. Using conditional statement for feature engineering
   Video: https://www.youtube.com/watch?v=Zs9NuYw-F7k
5. Summarizing data with Pivot Tables
   Video: https://www.youtube.com/watch?v=m0wI61ahfLc
6. Using Pivot Chart Analysis, Filters and Slicers
   Video: https://www.youtube.com/watch?v=mc7xO8F8Pj8
7. Compiling charts on a dashboard
   Video: https://www.youtube.com/watch?v=3oJxdvDerus

Project 2: Data Analytics & Visualization on Power BI

Project : **Dashboard in a Day**

About the project: 'Dashboard in a Day' is a comprehensive Power BI training course designed by Microsoft and presented by AINE AI as a Microsoft Partner Network member.

Difficulty Level: Advanced (Step-by-step guidance from scratch)

Tools to Use: Microsoft Power BI

Aim: Performing market share analysis and competitor analysis for Vanarsdel (the company) to take data-informed business decisions.

Objectives:

1. Create new measures, columns and tables as per the need of the analysis.
2. Create a Market Share Analysis dashboard to study the performance of the company with its competitors with visuals including
   a. Total Market Share metric for Vanarsdel – Card visual
   b. % Growth by different manufacturers – Stacked column chart
   c. Revenue by year and manufacturers – Line chart

    d. Revenue in different countries – Map chart
    e. Revenue and & growth by category, segment and year – Scatter chart
    f. Interactivity and filtering by a timeline - Slicer
3. Create a Competitor Analysis dashboard to understand and analyze the performance of Vanarsdel and its competitors in revenue generation
    a. Revenue vs Previous Year Sales – Gauge chart
    b. Animation to show changes in revenue over time – Play axis
    c. Revenue by Country for each competitor – Stacked bar chart
    d. Grouping variables to understand revenue by category and products – Matrix and conditional formatting
    e. Adding images as inputs in slicers
4. Customize the dashboard and each chart with various formatting features.

Dataset: Refer to the Data folder in this zipped file: DIAD by AINE AI

Skills to Use (Mandatory):

1. Connecting data from a variety of sources
2. Transforming data in Power Query Editor
3. Performing data transformation tasks such as column split, adding columns, renaming queries, changing data types, conditional statements, model analysis, writing DAX queries
4. Define business rules and KPIs
5. Explore data with different visualizations
6. Using filters on a visual, on a page and on multiple pages
7. Formatting visuals and dashboards
8. Build stunning interactive reports on multiple pages

Skills to Use (Optional):

1. Publishing and accessing the report on Power BI Service
2. Collaborating and distributing the report on various platforms
3. Implementing security and managing content at an enterprise level
4. Importing data in Power BI with live data sources and servers

Resources:

1. Refer to the User Manuals folder in this zipped file: DIAD by AINE AI
2. The 1st three lab manuals will help you achieve the objectives of the project utilizing the mandatory skills.
3. The last two lab manuals will help you use the optional skills.
4. Live doubt sessions will be held to help you with creating the dashboard and interpreting the observation.

5. Additional Resources (Videos):
    a. Power BI Essential Training (LinkedIn Learning): https://www.linkedin.com/learning/power-bi-essential-training-3/
    b. Splitting columns- https://www.youtube.com/watch?v=HQ0itUdOF-Q
    c. Replace values and change data types- https://youtu.be/UZQ9EFvEECY
    d. Creating a data table using DAX- https://youtu.be/gyvhM5eiT0U

Project 3: Data Reporting and Analysis with T-SQL

Project : **Product sales performance analysis using T-SQL**

About the project: Analysis of sales of various products by customers demographics and product categories for Adventure Works Cycles using T-SQL programming on Azure Data Studio.

Difficulty Level: Intermediate

Tools to Use: Azure Data Studio

Aim: Using T-SQL programming to summarize the sales of Adventure Works Cycles with respect to product characteristics, promotion cost and customer demographics.

Objectives:

1. Establish connection with SQL servers
2. Generate reports to containing details of the company's customers to support sales campaign
    a. Retrieve customer details
    Familiarize yourself with the Customer table by writing a Transact-SQL query that retrieves all columns for all customers.
    b. Retrieve customer name data
    Create a list of all customer contact names that includes the title, first name, middle name (if any), last name, and suffix (if any) of all customers.
    c. Retrieve customer names and phone numbers
    Each customer has an assigned salesperson. You must write a query to create a call sheet that lists:
    • The salesperson
    • A column named CustomerName that displays how the customer contact should be greeted (for example, "Mr Smith")
    • The customer's phone number.

3. Concatenating columns to create reports from same tables
    a. Retrieve a list of customer companies

You have been asked to provide a list of all customer companies in the format : - for example, 78: Preferred Bikes.

b. Retrieve a list of sales order revisions
The SalesLT.SalesOrderHeader table contains records of sales orders. You have been asked to retrieve data for a report that shows:
• The sales order number and revision number in the format () – for example SO71774 (2).
• The order date converted to ANSI standard format (yyyy.mm.dd – for example 2015.01.31).

4. Handling the NULL values in the database

Some records in the database include missing or unknown values that are returned as NULL. You must create some queries that handle these NULL fields appropriately.

a. Retrieve customer contact names with middle names if known
You have been asked to write a query that returns a list of customer names. The list must consist of a single field in the format (for example Keith Harris) if the middle name is unknown, or (for example Jane M. Gates) if a middle name is stored in the database.

b. Retrieve primary contact details
Customers may provide Adventure Works with an email address, a phone number, or both. If an email address is available, then it should be used as the primary contact method; if not, then the phone number should be used. You must write a query that returns a list of customer IDs in one column, and a second column named PrimaryContact that contains the email address if known, and otherwise the phone number.

c. Retrieve shipping status
You have been asked to create a query that returns a list of sales order IDs and order dates with a column named ShippingStatus that contains the text “Shipped” for orders with a known ship date, and “Awaiting Shipment” for orders with no ship date.

5. Querying tables to filter and sort data using

a. Retrieve a list of cities
Initially, you need to produce a list of all of your customers' locations. Write a Transact-SQL query that queries the Address table and retrieves all values for City and StateProvince, removing duplicates.

b. Retrieve the heaviest products
Transportation costs are increasing, and you need to identify the heaviest products. Retrieve the names of the top ten percent of products by weight.

c. Retrieve the heaviest 100 products not including the heaviest ten
The heaviest ten products are transported by a specialist carrier; therefore, you need to modify the previous query to list the heaviest 100 products not including the heaviest ten.

d. Retrieve product details for product model 1
Initially, you need to find the names, colors, and sizes of the products with a product model ID 1.

e. Filter products by color and size
Retrieve the product number and name of the products that have a color of 'black', 'red', or 'white' and a size of 'S' or 'M'.

f. Filter products by product number
Retrieve the product number, name, and list price of products whose product number begins 'BK-'.

6. Querying tables to join multiple tables and generate reports

a. Retrieve customer orders to generate invoice reports

As an initial step towards generating the invoice report, write a query that returns the company name from the SalesLT.Customer table, and the sales order ID and total due from the SalesLT.SalesOrderHeader table.

b. Retrieve customer orders with addresses
Extend your customer orders query to include the Main Office address for each customer, including the full street address, city, state or province, postal code, and country or region

c. Retrieve a list of all customers and their orders
The sales manager wants a list of all customer companies and their contacts (first name and last name), showing the sales order ID and total due for each order they have placed. Customers who have not placed any orders should be included at the bottom of the list with NULL values for the order ID and total due.

d. Retrieve a list of customers with no address
A sales employee has noticed that AdventureWorks does not have address information for all customers. You must write a query that returns a list of customer IDs, company names, contact names (first name and last name), and phone numbers for customers with no address stored in the database.

e. Retrieve a list of customers and products without orders
Some customers have never placed orders, and some products have never been ordered. Create a query that returns a column of customer IDs for customers who have never placed an order, and a column of product IDs for products that have never been ordered. Each row with a customer ID should have a NULL product ID (because the customer has never ordered a product) and each row with a product ID should have a NULL customer ID (because the product has never been ordered by a customer).

7. Working with conditions, aggregation and sub-queries in TSQL
Adventure Works products each have a standard cost price that indicates the cost of manufacturing the product, and a list price that indicates the recommended selling price for the product. This data is stored in the SalesLT.Product table. Whenever a product is ordered, the actual unit price at which it was sold is also recorded in the SalesLT.SalesOrderDetail table. You must use subqueries to compare the cost and list prices for each product with the unit prices charged in each sale.
    a. Retrieve products whose list price is higher than the average unit price
    Retrieve the product ID, name, and list price for each product where the list price is higher than the average unit price for all products that have been sold.
    b. Retrieve Products with a list price of $100 or more that have been sold for less than $100
    Retrieve the product ID, name, and list price for each product where the list price is $100 or more, and the product has been sold for less than $100.
    c. Retrieve the cost, list price, and average selling price for each product
    Retrieve the product ID, name, cost, and list price for each product along with the average unit price for which that product has been sold.
    d. Retrieve products that have an average selling price that is lower than the cost
    Filter your previous query to include only products where the cost price is higher than the average selling price

Dataset: Dataset can be obtained from AdventureWorks database by connecting Azure Data Studio with the below SQL server details:

- Server: sqlservercentralpublic.database.windows.net
- Database: AdventureWorks

- User: sqlfamily
- Password: sqlf@m1ly

Skills to Use (Mandatory):

1. Basics of relational databases and data schema
2. Querying tables with SELECT statement
3. Working with data types and type casting
4. Exploring and filtering data containing NULL values
5. Removing duplicates
6. Sorting Results
7. Filtering and using predicates
8. Joins in databases
9. Working with subqueries
10. Working with UNION, INTERSECT and EXCEPT to create reports from multiple tables.

Skills to Use (Optional):

In this project, you will work on a database provided to you via SQL Server. The database is read-only and hence, you won't be able to make any changes to it. However, you can also establish your personal local SQL server and maintain tables into it. The below two tutorials are good to go through to understand the setting up of a personal SQL server and creating your own database. These skills are not required for the given project.

1. Creating user account in SQL Server (need to install SQL Server)
   Video: https://www.youtube.com/watch?v=11Rx35l8Khc
2. Create, alter and drop table in SQL Server
   Video: https://www.youtube.com/watch?v=hCiwBl-kb4g

Resources:

1. What is SQL? https://www.linkedin.com/learning/sql-data-reporting-and-analysis-2/what-is-sql
2. Read the data in a table using SELECT, FROM and WHERE: https://www.linkedin.com/learning/sql-data-reporting-and-analysis-2/retrieve-data-with-select
3. Filtering data with WHERE clause and Logical operators: https://www.linkedin.com/learning/sql-data-reporting-and-analysis-2/filter-results-with-where-clause

4. Logical operators in T-SQL: https://docs.microsoft.com/en-us/sql/t-sql/language-elements/logical-operators-transact-sql
5. Use LIKE, IN, and wildcards with WHERE: https://www.linkedin.com/learning/sql-data-reporting-and-analysis-2/use-like-in-and-wildcards-with-where
6. Sorting data with ORDER BY: https://www.linkedin.com/learning/sql-data-reporting-and-analysis-2/sort-sql-results-with-order-by
7. String functions - CONCAT, LENGTH, UPPER, LOWER, RIGHT, LEFT: https://www.linkedin.com/learning/sql-data-reporting-and-analysis-2/use-string-functions-on-your-data
8. Giving new column names in the result columns: https://www.linkedin.com/learning/sql-data-reporting-and-analysis-2/change-report-headings-with-alias
9. Grouping data with count: https://www.linkedin.com/learning/sql-data-reporting-and-analysis-2/use-group-by-with-count
10. Working with data types and type casting : https://youtu.be/5MUbbiMSLQg
11. Removing duplicates : https://youtu.be/g9LjsBMvW28
12. Joins in databases : https://www.youtube.com/watch?v=zGSv0VaOtR0
13. UNION & UNION ALL in databases: https://www.linkedin.com/learning/sql-data-reporting-and-analysis-2/combine-rows-with-sql-union
14. UNION, INTERSECT & EXCEPT in databases: https://youtu.be/6M8dGCuSHT0
15. Using subqueries in databases: https://youtu.be/GpC0XyiJPEo

Project 4: Data Visualization and Storytelling using Tableau

Project : **A timeline study of Covid-19 cases globally and the role of demographic, economic and public health factors in helping the pandemic spread or reduce.**

About the project: This project involves comprehensive study of Covid-19 with respect to its timeline as well as trying to explore the demographic, economic and public health situations in different countries that have helped in the spread or reduction. It also involves a detailed analysis of the impact of pandemic in your country.

Difficulty Level: Beginner, Intermediate and Advanced (Please look at Objectives)

Tools to Use: Tableau Public or Tableau Desktop (Academic Edition)

Create Profile and downloading Tableau Public:
https://public.tableau.com/en-us/s/

Downloading Tableau Desktop Academic:
https://www.tableau.com/academic/students

Those using Tableau Desktop Academic version would need a license key to use the tool. Please use this link to submit your details to verify you are a full-time student:
https://www.tableau.com/academic/students#form

Aim: Create a data story explaining the timeline of the Covid-19 pandemic in different countries. Explore the demographic, economic and public health situations in the countries which are worst or least hit. Prepare a detailed analysis of the Covid-19 timeline and its impact in your country, along with the factors that may have contributed to its growth or decline.

Objectives:

1. [All difficulty levels] Explore the data after loading and do basic data preparation, if required.
2. [All difficulty levels] Create an animated time series of Covid-19 active/death cases in different countries with an animation that changes with dates.
   Note: Play button is not available in Tableau Public. However, you can manually play it to show the trend month-wise.
3. [Intermediate level] Create a dashboard dedicated to the Covid-19 situation in your country. It can contain the metrics like total cases, deaths, vaccinations, hospital admissions etc.
4. [Intermediate level] Create a Covid-19 Vaccination Dashboard to show the status of vaccination globally.
5. [Advanced level] Create dashboards to explore various variables related to Covid-19 impact globally as given in the data. You can use the column which you think should be used in your analysis.
6. [Advanced level] Create dashboard(s) to show how demographic, economic and public health status have helped Covid-19 to grow or decline globally.
7. Storyboard creation is required for those who are working on intermediate and advanced level tasks.
8. Publishing your visual/dashboard/storyboard on Tableau Public.

Dataset: The number of Covid-19 cases are continuously changing and hence, you should also use the most updated data for your analysis. Please go to this link and scroll down to download the most recent dataset in Excel or CSV format:

https://github.com/owid/covid-19-data/tree/master/public/data/

Skills to Use (Mandatory):

1. Connecting a data source in Tableau Public
2. Data exploration and changing the data type of a column
3. Creating visuals of different variety to explore different combination variables
4. Using filters, sorting and grouping of and visuals
5. Using Page Play animation
6. Creating a dashboard containing multiple visuals. [NOT mandatory for beginners]
7. Creating storyboards using multiple dashboards
8. Publishing the dashboard on Tableau Public

Skills to Use (Optional):

1. Connecting pages with buttons
2. Using URLs and images
3. Using calculations to create new measures or dimensions

Resources:

Tableau maintains a large community to help its users find answers of their product or feature related queries. Since this project would involve a wide number of Tableau features, you are suggested to first explore the answers on Tableau Community.

Additionally, please login to your Tableau Public account to access Free Training Videos by Tableau on various topics. Please select the Tableau version you are using before accessing the video content.

Project 5: Essentials of Python Programming

Project: **Developing 'Rock, Paper and Scissors' game using Python programming**

About the project: The project leverages the basic concepts of Python programming such as variables, data types, loop and conditional statements to develop the game of Rock, Papers and Scissors.

Difficulty Level: Beginner

Tools to Use: Any Python IDE such as Jupyter Notebook, Azure Data Studio, Google Colaboratory or PyCharm. This project will guide you with working on Jupyter

Notebook (Anaconda 3). Please click here and follow the link to download and install the tools.

Aim: Write a Python program to develop a Rock, Papers and Scissors game to be played against a computer.

Objectives:

1. Understand the environment of the Jupyter Notebook on Anaconda 3.
2. Using different cells and comments in the notebook to mention explanations of your program and syntaxes.
3. Understanding basic data types and lists in Python.
4. Working with lists and accessing its elements.
5. Taking user inputs and printing them.
6. Working on the random library
7. Initializing the value of a variable
8. Creating loops asking the play the game multiple times unless the user enters Q
9. Using conditional operators to play the game with different options in different games.

Dataset: This Project does not require any specific dataset. You can create your own variables and values to develop the game. In the list that you create, mention these elements must be there: Rock, Paper, and Scissors.

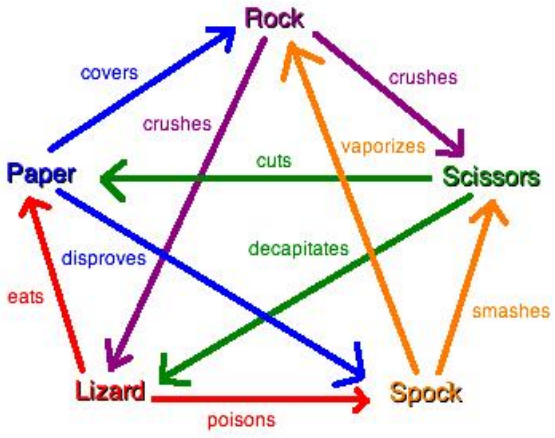
Skills to Use (Mandatory):

1. Using Python 3 in Jupyter Notebook
2. Using different types of cells – Code, Markdown and Heading
3. Adding comments in the code cells
4. Creating variables with basic data types
5. Creating list variables
6. Taking inputs from the users
7. While loop
8. Using comparison operator to check the inputs
9. Conditional statements
10. Nested conditional statements in Python
11. Break statement
12. Concatenated print statements

Skills to Use (Optional):

1. Using For loop to let the user opt for how many rounds he/she wants to play the game.

2. Formatting text inputs from the user so that the program does not show inputs invalid if the inputs are not case-sensitive.
3. Use isdigit() to mark the input invalid if it is a digit.
4. Add two more elements in the game – Lizard and Spock. The game structure should look like:



Resources:

1. Understanding Jupyter Notebook and its interface
   Article: How to use Jupyter Notebook – by Codecademy
   Video: Jupyter Notebook System – by Coursera
2. Learning Python (for Beginners) - By LinkedIn Learning: https://www.linkedin.com/learning/learning-python
3. Comments in Python : https://youtu.be/lwmAVBMEzkM
4. Variables – Integers: https://youtu.be/rHjAiQDw9V8
5. Variables – Strings: https://youtu.be/36AClUO0Yi8
6. User Input: https://youtu.be/YLS7-ZY9HOE
7. Formatting String Inputs: https://youtu.be/dxKQAdsVg4o
8. Comma Separated String Printing: https://youtu.be/5O8-tKqHI5I
9. Comparison operators: https://youtu.be/ikB7_rlRD6w
10. Conditional statements: https://youtu.be/Ecm6La33E1k
11. Nested conditional: https://youtu.be/hadVXHyBKfU
12. While loop: https://youtu.be/F2Ihn9b1_OM
13. Conditional Boolean strings (isdigit(), istitle()): https://youtu.be/RWzmEWp4dII
14. Break() and Continue() in a WHILE loop: https://www.youtube.com/watch?v=BTaPo33TBIM
15. Variables – Knowing the Data Types: https://youtu.be/KJBdE6u-Uho
16. Addition – Numbers & Strings: https://youtu.be/BbkG03VKoWo
17. Variable Addition: https://youtu.be/HpuQrNUp38s
18. Creating a list in Python: https://youtu.be/bRtbAg5v2TM
19. Accessing the elements of a list: https://youtu.be/QkOftrcxtCc
20. Adding element into a list: https://youtu.be/dGTNjOOqEcI

21. Delete specific items from a list: https://youtu.be/OKw9y3MgBfY

Project 6: Data Manipulation and Visualization using Python

Project : **Creating visuals and charts using Python libraries**

About the project: The project contains three parts:

1. Data manipulation using Pandas (Hands-on training)
2. Data visualization using Matplotlib (Hand-on training)
3. Data manipulation and analysis on Python (Guided Assessment)

Difficulty Level: Intermediate

Tools to Use: Google Colaboratory

Aim: Performing data manipulation and visualization techniques using Python to clean, transform and visualize different datasets.

Objectives:

1. Creating, exploring and modifying of BRICS data using Pandas library
2. Creating customized visualizations using Matplotlib library
3. Performing data manipulation such as cleaning data, adding columns, sorting data etc using Python programming

Skills to Use:

1. Writing Python codes on Google Colaboratory
2. Accessing data from different sources
3. Creating dataframes, setting index and transforming the data
4. Generating reports by accessing elements of the dataframes
5. Developing visuals and charts using Matplotlib library
6. Customizing and analyzing charts in Python

Resources & Datasets:

Since the entire project is a guided training and assessment, the resources, datasets as well as starter codes – all will be mentioned in this Python Notebook at their respective position:
https://colab.research.google.com/drive/1B7303PgWbef5cdnlQM6fQU3zE5RaYWaZ?

Project 7: Statistical Analysis and Hypothesis Testing

Project : **Increasing YoY revenue from game purchases by increasing retention rate of gamers**

About the project: The project involves working on data related to Cookie Cats – a hugely popular puzzle game. As players progress through the levels of the game, they will occasionally encounter gates that force them to wait a non-trivial amount of time or make an in-app purchase to progress. In addition to driving in-app purchases, these gates serve the important purpose of giving players an enforced break from playing the game, hopefully resulting in the player's enjoyment of the game being increased and prolonged. But where should the gates be placed and how the placement of the gates can retain the players for more time.

The project requires you to:

1. Perform exploratory data analysis of the given datasets and generate their statistical summary.
2. Perform A/B testing between the two groups of data to analyze the impact on player retention if the first gate in Cookie Cats is moved from level 30 to level 40.

Difficulty Level: Intermediate

Tools to Use: Python (preferred) or R Programming

Aim: Even though the overall subscription for the game is growing, the revenue from in-game purchases has been declining and many players are uninstalling the game after playing for a few days. What recommendations can you suggest increasing the in-game purchase and retaining the players?

Objectives/Exercise:

The overall objective of the project is to test the company's CEO's hypothesis that moving the first gate from level 30 to level 40 increases retention rate and the number of game rounds played. The CEO believes that players are churning because the first gate encountered at level 30 is too early which forces players to wait before they can proceed further in the game.

In order to increase player retention rate, developers ran AB-test by moving the first gate from level 30 to level 40 for some players i.e.,

- group A would encounter the gate at level 30, and
- group B would encounter the gate at level 40

To achieve the overall objectives, this work plan can help:

1. Perform initial data preparation.
2. Generate statistical summary and plot charts to answer:
    a. What is the overall 7-day retention rate of the game?
    b. How many players never played the game after installing it?
    c. Does the number of players decrease as the levels become difficult?
3. Generate crosstab for two player groups to understand the difference in the 1-day and 7-days retention rate and total number of game rounds played.
4. Perform two-sample test for groups A and B to test statistical significance amongst the groups in the sum of game rounds played. Here, you can:
    - Check the assumptions of two sample test:
        - i. Normal distribution – Apply Shapiro test
        - ii. Homogeneity of variance – Apply Levene's test
    - Apply the relevant two sample significance test method based on the results from the tests for normality and homogeneity
5. Analyze the significance of the test results and decide which level has more advantage in terms of player retention?
6. Use bootstrap resampling to plot retention rate distribution for both groups to visualize the effect of different versions of the game on retention.

Materials:

Dataset: https://drive.google.com/file/d/1ZhUjUuG9J53g-rtEtHVucvfwSkgd3W5G/view?usp=sharing

Starter code: *https://drive.google.com/file/d/1UJ5Djmejz-tzSmoihibc8UyWOa2xbHBW/view?usp=sharing*

Skills/Concepts to Use:

1. Slicing and dicing data frames to generate summary statistics
    a. Filtering data frames based on column(s)

b. Splitting data into multiple data frame(s)

c. Shape of data frame, counting rows and length

d. Adding/deleting column(s) to a data frame

e. Aggregations and merging data frames

f. Grouping, pivot table, cross tabs

g. Scatter plots, histogram, box plots, etc.

2. Hypothesis testing

    a. Null and alternate hypothesis

    b. P-value and significant testing based on confidence intervals

3. One sample t-test - One-tailed and two-tailed

4. Two sample t-test

    a. Assumptions –normality test and homogeneity of variances

    b. Types of two sample t-test

        i. z-test

        ii. t-test with equal variances

        iii. t-test with unequal variances

Resources:

1. Essential Statistics Training on Excel (LinkedIn Learning): https://www.linkedin.com/learning/excel-statistics-essential-training-1-2
2. Essential Math for Machine Learning: Python Edition https://www.linkedin.com/learning/essential-math-for-machine-learning-python-edition
3. Descriptive statistics:

    a. Individuals and variables: https://youtu.be/noV7b2mZ6VU

    b. Numerical and categorical data: https://youtu.be/FsGerboj9Sc

    c. Types of data: https://youtu.be/ga3cNZ7ZcoQ

d. Using histograms to summarize data: https://youtu.be/QCvvbck8B-A

e. Calculating Mean, Median and Mode: https://youtu.be/Ciizn_yX46s

f. Calculating Range, Variance and Standard Deviation: https://youtu.be/X6i5orYSU5M

g. Skewness analysis: https://youtu.be/lBN0Q2W-4ec

h. Boxplot analysis: https://youtu.be/9ftg9uD5qTE

2. Sampling and Confidence Intervals:

a. Sampling and sampling distributions: https://youtu.be/f87kb6kBPPk

b. Confidence intervals: https://youtu.be/EqbAQxQiQ00

c. Confidence intervals examples: https://youtu.be/YkZSwRq2EcA

d. Population and samples: https://youtu.be/K6kEAV-poy0

e. Sampling strategies: https://youtu.be/PohcgP-8m3M

f. Problems in sampling: https://youtu.be/XPPMLWeDigA

g. Mean, variance and standard deviation of sample mean: https://youtu.be/zvtThcEWt6s

h. Estimating population proportion using samples: https://youtu.be/boSPWIUniqs

i. Sample size determination: https://youtu.be/1im1VCSnkYw

3. Hypothesis testing:

a. Overview of hypothesis testing: https://youtu.be/Nr501ePJ-ow

b. Null and alternate hypothesis: https://youtu.be/hPyjpQ_rhds

c. One and two tailed tests: https://youtu.be/wNFJ1ICPE-Y

d. Choosing between one-tailed and two-tailed test: https://youtu.be/Dz2wuxvh0h8

e. Type I and type II error: https://youtu.be/ca-GMlTZsnM

f. Critical region: https://youtu.be/ni4r8tg8F3g

g. One sample Z-test: https://youtu.be/Y0z6FMTmBAc

h. P-values: https://youtu.be/jzSCUtYmz2Q

i. T random variable: https://youtu.be/70qpGzIrPSQ

j. One sample T-test : one-tailed: https://youtu.be/43-vHhD2Yro

k. One sample T-test: two-tailed: https://youtu.be/VYfk0ohvafM

l. Single Sample test for population proportion: https://youtu.be/OiJXxVX4r7Q

m. Testing equity of variances: https://youtu.be/NCR-R1GPsvk

4. Sample testing:

a. Four types of tests: https://youtu.be/JWBGbxxlumw

b. Which type of the four tests to use: https://youtu.be/O2gte4GP49A

c. Two sample Z-test: https://youtu.be/A_0Pe9iwloA

d. Equal variance T-test: https://youtu.be/Uebs6myfSGU

e. Unequal variance T-test: https://youtu.be/_FoPGMQxOJ0

f. Idea of pairing: https://youtu.be/cv7wUungIcA

g. T-test paired two sample: https://youtu.be/VnGeVz4xi8E

5. Contingency Table and Hypothesis of Independence:

a. Introduction to contingency table and hypothesis of independence: https://youtu.be/yje5XX97gQA

b. Chi-squared statistics: https://youtu.be/kdteJPdNsBM

c. Computing Chi-squared statistics: https://youtu.be/5D3p2NDahWs

d. Conducting the hypothesis test and computing p-value: https://youtu.be/ony4e0et9Uk

Project 8: Machine Learning for Predictive Analytics

Project : **Reducing monthly churn by identifying high risk customers well in advance**

About the Project: The project relates to applying predictive analytics on customer churn. A major telecom company's postpaid business of voice-only plans is struggling to maintain its strong foothold in local market because of:

- High churn rate amongst customers leading to a revenue decline of ~500k USD every month
- Decline in overall customer base (high churn rate combined with low acquisition rate), leading to a decline in total market share

Aim:

1. Build a classification model to predict churners one month in advance
2. Identify key churn drivers

Hypothesis: Company CEO believes that existing models can predict churners precisely, but it's too late to take any retention actions, as customer usage has significantly declined by then.

Objectives/Exercise:

1. Perform initial data preparation
    a. Number of customers with zero monthly revenue?
    b. Number of customers with missing values percentage > 5%?
    c. Remove outliers for columns 'UniqueSubs' and 'DirectorAssistedCalls' is any.
2. Perform exploratory analysis to analyse customer churn
    a. Do customers with high overage minutes also have high revenue?
    b. Does high number of active subscribers lead to low monthly revenue?
    c. Does credit rating have an impact on churn rate?
3. Create additional features to help predict churn
    a. Percent of current active subs over total subs
    b. Percent of recurrent charge to monthly charge
    c. Percent of overage minutes over total monthly minutes
4. Build classification model to predict customer churn

a. Build a simple logistic regression model to predict churn and evaluate model accuracy on test data set
b. Build Random Forest classifier to compare model accuracy over the logistic regression model
c. Identify most important features impacting churn (Model evaluation metrics to be used: GINI, AUC, Precision and Recall)

5. Use the 'Prediction Data' provided to predict churners using the best model identified in step 4
6. Calculate lift chart and total monthly revenue saved by targeting top 10-20% of the customers using your best predictive model

Materials:

Dataset:
https://drive.google.com/file/d/1ikpaSxH_O5tFsgUUgB1M04VM65HG3Qx9/view?usp=sharing

The folder contains two .csv files. You have to work on 'Telecom Data' for the first 4 questions/objectives. For prediction in objective# 5, you should use the 'Telecom - Prediction Data' file.

Starter Code:
https://drive.google.com/file/d/1Rb8cOhowc2AOEnZAynleouwdRxpjfEKQ/view?usp=sharing

Data Dictionary

Columns	Description
CustomerID	Unique Customer ID
MonthlyRevenue	Monthly revenue in USD
MonthlyMinutes	Monthly minutes
TotalRecurringCharge	Recurring charges in the past month
DirectorAssistedCalls	Automated calll (directory assisted calls)
OverageMinutes	Extra minutes above the postpaid allocation
RoamingCalls	Minutes on calls while roaming
PercChangeMinutes	percentage change in minutes from previous month
PercChangeRevenues	percentage change in revenue from previous month
DroppedCalls	Dropped calls in minutes
BlockedCalls	Blocked calls in minutes
UnansweredCalls	Unanswered calls in minutes
CustomerCareCalls	Customer care call duration in minutes
ThreewayCalls	Minutes spend on Conference calls
ReceivedCalls	Total received calls in minutes
OutboundCalls	Marketing calls received from customer service in minutes
InboundCalls	total duration in minutes of calls made to customer service
PeakCallsInOut	Incoming/outgoing calls during peak time
OffPeakCallsInOut	Incoming/outgoing calls during off peak time
DroppedBlockedCalls	Summation of dropped and blocked

Columns	Description
CallForwardingCalls	Minutes in call forwarding
CallWaitingCalls	Minutes spend on hold during call
MonthsInService	Total months in service
UniqueSubs	total number of unique sim card subscriptions (includes inactive connection)
ActiveSubs	total # of active subscriptions
ServiceArea	Service area of the subscription
Handsets	Total # of handsets
HandsetModels	total # of unique hadnset model
CurrentEquipmentDays	Number of days since the activation of current equipment
AgeHH1	Primary holder
AgeHH2	Secondary holder
ChildrenInHH	Flag indicating childen in household
HandsetRefurbished	Handset refurbished flag (returned to company and then they sell it to different customer)
HandsetWebCapable	Internet connectivity
TruckOwner	Flag indicating whether the Customer owns a truck
RVOwner	Customer owns RV or not
Homeownership	Home owned by customer
BuysViaMailOrder	Whether customer has bought anything via clicking an option on email
RespondsToMailOffers	Flag indicating whether customer responds to mail offers
OptOutMailings	Whether the customer has opted out of mailing

Columns	Description
NonUSTravel	Flag indicating whether the customer has travelled outside of US
OwnsComputer	whether customer Owns computers
HasCreditCard	Whether customer owns a credit card
RetentionCalls	Whether customer responded to retention calls
RetentionOffersAccepted	Whether customer accepted retention offers
NewCellphoneUser	New cellphone user flag
NotNewCellphoneUser	Total not new cellphone user flag
ReferralsMadeBySubscriber	Total referrals made by subscriber
IncomeGroup	Income group
OwnsMotorcycle	Owns motorcycle flag
AdjustmentsToCreditRating	Number of time the credit rating has ranged in past 1 year
HandsetPrice	Price of the handset in USD
MadeCallToRetentionTeam	Flag indicating whether customer made call to retention team
CreditRating	Credit rating of cthe customer
PrizmCode	Area group of customer home location
Occupation	Type of occupation
MaritalStatus	Marital status

Skills to use:

1. Exploratory data analysis
    a. Missing value identification and treatment, outlier detection, etc.

    b. Univariate analysis - histograms to check distribution, box/violin plot, summary stats such as mean, median, mode, etc.
    c. Correlation analysis, scatter plots
    d. Time analysis to monitor trend and seasonality
2. Feature engineering
    a. Target variable creation (dependent variable) - align with the objectives
    b. New feature creation based on business and the objectives
    c. Trend variables such as moving average, st. dev. over time, etc.
3. ML model build
    a. Model type – classification vs regression
    b. Algorithm type – logistic regression, linear regression, Random forest, etc.
    c. Train and test data split (e.g., 70:30 split)
    d. Hyperparameter tuning for model optimization
    e. Feature importance
4. Model evaluation
    a. Model evaluation metrics
    b. Regression – R-Square, Mean Absolute Error (MAE), Mean Square Error
    c. Classification – Confusion metrics, precision, recall, GINI, AUC, etc.
    d. Lift and Gain charts
    e. Prediction data set for model validation

Project 9: Image Classification using Deep Learning

Project : **Classifying Covid-19 positive and negative patients from X-ray images**

About the Project: The outbreak of COVID-19 has had an immense impact on world health and daily life in many countries. The first imaging procedure that played an important role in COVID-19 treatment was the chest X-ray. Radiological imaging is often used as a method that emphasizes the performance of chest X-rays. Recent findings indicate the presence of COVID-19 infections in the patients with irregular findings on chest X-rays. There are many reports on this topic that include machine learning strategies for the identification of COVID-19 using chest X-rays.

This project uses radiological imaging to determine whether the scanned patient has COVID-19 or not.

Aim: With the Chest X - Ray dataset, develop a Deep Learning Model to classify the X Rays of Healthy vs Corona positive patients.

Objectives/Exercises:

1. Import the dataset in python Notebook
2. Explore the dataframe
3. Perform data transformation to preprocess the images to convert the images to the same size and greyscale.
4. Perform normalization techniques on the images
5. Split the dataset into training and testing sets.
6. Create a Convolution Neural Network (CNN) model to classify the images into positive and negative COVID-19 infections.
7. Test the CNN model and critically evaluate the performance of the model

Tools & libraries to Use: JupyterNotebook or Google Colab, Python3, Numpy, OpenCV, Tensorflow, Keras etc

Datasets:

Covid Positive Images: https://github.com/ieee8023/covid-chestxray-dataset/tree/master/images

Covid Negative (Normal): https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia