COMM 337 Group Project 2

Instructions

The main objective of this project is to motivate you to dive into real-world data and use Python to draw insights from it. This is a flexible and open-ended assignment, so you and your group will have the freedom to choose a dataset and conduct a comprehensive analysis to generate valuable business insights. There are various potential directions you can take in your analysis, and these may include, but are not limited to:

- drawing creative insights and useful knowledge through descriptive analysis and visualization of the data,
- identifying patterns or relationships in data with unsupervised machine learning algorithms,
- making predictions about potential outcomes by training supervised machine learning models.

You can choose any dataset provided in this folder (link), or you can gather your own dataset from any open-source data website or through web scraping.

Evaluation Criteria

1. Research Proposal Presentation [40%]

Each group needs to prepare a **3-minute** presentation describing their research plans. The objective of this presentation is to learn from and also inspire other classmates on how to dig out the full potential of each dataset. Possible topics include: (i) selected data; (ii) research questions; (iii) approaches to answer the questions; (iv) expected results; (v) preliminary analysis.

Specific requirements for this presentation are listed below:

- **At least one** group member needs to present the proposal in class on April 13. We will strictly control each presentation to **3 minutes only**, so it is strongly recommended to limit your slides to **no more than 5 pages**.

- Slides for the presentation need to be submitted **in PDF format** to "**Project2-Presentation**" on Canvas before **April 12, 11:59 pm**.

- The presentation will be evaluated by peers. Each student will be randomly assigned to 3 presentation evaluations on Canvas. Please submit your comment and score on Canvas before **April 13, 11:59 pm**.

2. *Data Analysis Report [60%]*

Conduct a thorough analysis using the data and document your work in a Jupyter Notebook. The notebook should include:

- Information on group members (name + student ID) and work distribution.
- An introduction to the research question and dataset.
- Detailed analysis of the question. You need to include your code and display the results (e.g., data visualization, summary statistics, machine learning predictions, etc.) in the notebook.
- Insights and conclusions.

If you want to refer to some code from other resources, please cite them in your notebook as before. **We will conduct a plagiarism check for all notebooks. If we find evidence of plagiarism, we will directly report it to the UGO.**

Submit your Jupyter Notebook to "**Project2-Report**" on Canvas by **April 17, 11:59 pm**. If you use a different dataset, please also include it in the submission so that we can replicate your analysis.

Important Dates

April 12, 11:59 pm	Submit your presentation slides (PDF format) on Canvas.
April 13	Proposal presentation in class. Submit your comments and grades for the peer evaluation.
April 17, 11:59 pm	Submit full report (Jupyter notebook format) on Canvas.

Datasets

1. Netflix Movies and TV Shows

The dataset comprises a record of approximately 6,000 movies and TV shows that are currently accessible on Netflix, along with the staff members such as directors and actors who contributed to these productions. The metadata for movies and shows is gathered from Kaggle (https://www.kaggle.com/datasets/dgoenrique/netflix-movies-and-tv-shows), and the crew data is collected from IMDb (https://www.imdb.com/interfaces/). The data files are named "*titles.csv*" and "*crew_info.csv*". You can merge these two files with "*imdb_id*".

Data Dictionary

titles.csv

Column	Meaning
id	The title ID on JustWatch.
title	The name of the movie or show.
show_type	TV show or movie.
description	A brief description.
release_year	The release year of the show or movie.
age_certification	The age certification.
runtime	The duration of the movie or TV show in minutes. For TV shows, this field indicates the average duration of each episode.
genres	A list of genres.
production_countries	A list of countries that - produced the show or movie.
seasons	Number of seasons if it's a SHOW
imdb_id	The movie or TV show ID on IMDB.
imdb_score	Score on IMDB.
imdb_votes	Votes on IMDB.
tmdb_popularity	Popularity on TMDB.
tmdb_score	Score on TMDB.

crew_info.csv

Column	Meaning
imdb_id	The movie or TV show ID on IMDB. Use this to match with title.csv.
ordering	A number to uniquely identify rows for a given imdb_id.
name_id	Unique id for the crew member.
category	The category of job that person was in.
job	The specific job title if applicable, else '\N'.
characters	The name of the character played if applicable, else '\N'.
primaryName	The real name of the crew member.
birthYear	In YYYY format.
deathYear	In YYYY format if applicable, else '\N'.
primaryProfession	The top-3 professions of the person.

2. British Columbia Car Crashes (2017-2021)

This dataset contains the car crash statistics in British Columbia provided by ICBC from 2017 to 2021. More details of this dataset can be found in the ICBC public data website (https://public.tableau.com/app/profile/icbc/viz/ICBCReportedCrashes/ICBCReportedCrashes). The data file is named "*BC_car_crash.csv*".

Data Dictionary

Column	Meaning
Year	Year of crash.
Month Of Year	Month of crash (e.g., January).
Day Of Week	Day of week of crash (e.g., Monday).
Time Category	Time of crash in three-hour groupings (e.g., 15:00-17:59).
Crash Severity	The level of crash severity: • Casualty crash: a crash resulting in an injury or fatality. • Property damage only crash: a crash resulting in material damages to properties (vehicle or non-vehicle) with no injuries or fatalities.
Derived Crash Configuration	• Head on • Multiple impacts • Overtaking • Rear end • Rear to rear • Side Impact • Side swipe - opposite direction • Side swipe - same direction • Single vehicle • Undetermined • Conflicted (where there are multiple/different configurations reported)
Intersection Crash	Yes if crash occurred at an intersection between roads. Includes interchanges, onramps and off-ramps.
Heavy Veh Flag	Yes if a vehicle involved had gross vehicle weight of more than 10,900 kg.
Animal Flag	Yes if one or more animal was involved.
Cyclist Flag	Yes if one or more cyclist was involved.
Motorcycle Flag	Yes if one or more motorcycle was involved. Motorcycles include mopeds, limited speed motorcycles, scooters and trikes.
Parked Vehicle Flag	Yes if one or more vehicles involved was parked at the time.
Parking Lot Flag	Yes if the crash occurred in a parking lot or with a parked vehicle.
Pedestrian Flag	Yes if one or more pedestrian was involved.
Region	The ICBC region in B.C. in which the crash occurred: • Lower Mainland

	• North Central • Southern Interior • Vancouver Island • Unknown
Municipality Name	The municipality in which the crash occurred (e.g., Vancouver), based on claim reports.
Street Full Name	The name of the street, thoroughfare, or road infrastructure component where the crash occurred.
Road Location Description	The description of the intersection, road segment or road infrastructure where the crash occurred.
Total Crashes	The distinct number of crashes with the above parameters.
Total Victims	The total number of victims injured or killed in crashes with the above parameters.

3. British Columbia Driver License Test Results (2017-2022 June)

This dataset contains the results for all driver road tests in British Columbia provided by ICBC from 2017 to 2022 June. More details of this dataset can be found in the ICBC public data website (https://public.tableau.com/app/profile/icbc/viz/Roadtests/RoadTests). The data file is named "*BC_driver_license_test.csv*".

Data Dictionary

Column	Meaning
test_id	Unified test id.
driver_exam_breakdown	The specific class for the exam.
year	Year of the test.
exam_month	Month of the test.
exam_reason	The reason that the exam was taken.
exam_result	The results of an exam.
exam_type	Type of exam taken.
gender	Driver's or examinee's gender. Due to system limitations, gender identities other than male or female are included as "Information not available".
office_municiplaity	The municipality of ICBC driver license office (e.g. Vancouver)
region	The ICBC region in B.C.: • Lower Mainland • North Central • Southern Interior • Vancouver Island Unknown
exam_class	The specific class for the exam.
exam_date	Date of the test.

4. Stroke Data

This data provides participants' health statistics and whether they suffered from a stroke. The data file is named "*healthcare-dataset-stroke-data.csv*". The dataset is downloaded from Kaggle (https://www.kaggle.com/datasets/thedevastator/bigmart-product-sales-factors).

Data Dictionary

Column	Meaning
id	Unified participant id.
gender	Male, female, other
age	The age of the participant.
hypertension	Whether the participant has hypertension.
heart_disease	Whether the participant has heart diseases.
ever_married	Whether the participant is ever married.
work_type	Private, self-employed, and other.
residence_type	Urban or rural.
avg_glucose_level	The average glucose level of the participant.
bmi	Body Mass Index of the participant.
smoking_status	Whether the participant smokes.
stroke	Whether the participant suffered from a stroke.

5. BigMart Product Sales

This dataset contains the sales data from BigMart, including basic product attributes and their final sale. The data file is named "*bigmart_product_sales.csv*". The dataset is downloaded from Kaggle (https://www.kaggle.com/datasets/thedevastator/bigmart-product-sales-factors).

Data Dictionary

Column	Meaning
Item_identifier	Product id.
Item_weight	Weight of the product in kilograms. (Numeric)
Item_Fat_Content	The fat content of the product. (Categorical)
Item_Visibility	The visibility of the product in store or online. (Numeric)
Item_Type	The type of product, such as limited offers or no offer. (Categorical)
Item_MRP	The maximum retail price of the product. (Numeric)
Outlet_Establishment_Year	The year the outlet was established. (Numeric)
Outlet_Size	The size of the outlet, either retail or supermarket. (Categorical)
Outlet_Location_Type	The type of location of the outlet, such as urban or rural area. (Categorical)
Outlet_Type	The type of outlet, such as sales departmental store or supermarket. (Categorical)
Item_Outlet_Sales	The sales of the product in the outlet. (Numeric)

6. Kickstarter Projects

This dataset contains around 370K projects from the famous crowdfunding platform Kickstarter. The data file is named "*kickstarter_projects.csv*". The data is collected from Kaggle (https://www.kaggle.com/datasets/ulrikthygepedersen/kickstarter-projects).

Data Dictionary

Column	Meaning
ID	Project id.
Name	Project Name.
Category	Category of the project (e.g., film &video, music).
Subcategory	Subcategory of the project (e.g., documentary, product design).
Country	Country where the project is launched.
Launched	Time of the project launch.
Deadline	The deadline for crowdfunding.
Goal	Amount of money the creator needs to complete the project (USD).
Pledged	Amount of money pledged to by the crowd (USD).
Backers	Number of backers who pledged to the project.
State	Current condition the project is in (as of 2018-01-02).

7. Others

Feel free to use any other datasets you are interested in! Please include them in the submission so that we can replicate your analysis and results.