Data Science
Foundation Projects

Project Guidelines

Project has to be submitted in **video format.**

These are the guidelines to be followed while making the video.

1. Prepare the report in form of a power point presentation. Below are the guidelines for the report.
    a. Team info
    b. Domain & topic of project
    c. Introduction (brief info on project)
    d. Dataset description
    e. Business questions identified (at least 7-8 questions)
    General format:
    Question 1
    Approach
    Findings & Visualizations

2. Record your presentation using applications like **HYFY** Google chrome extension.

3. Convert the video into mp4 format.

4. Video length should be between 10-15 mins.

Top three video presentations will be featured on our **Youtube channel**

Domain: Airlines
Project 01: Analyze NYC-Flight data

This dataset contains information about all flights that departed from NYC (e.g. EWR, JFK and LGA) in 2013: 336,776 flights in total. The following are the types of question you can ask:

Variable description:

Name	Description
year	2013
month	1-12
day	Day of the month (1-31)
dep_time	Departure times, local timezone
sched_dep_time	Scheduled departure time
dep_delay	Departure delay, in minutes, Negative times represent early departures
arr_time	Arrival times, local timezone
sched_arr-time	Scheduled departure time
arr_delay	Arrival delay, in minutes, Negative times represent early arrivals
carrier	Two letter carrier abbreviation
flight	Flight number
tailnum	Plane tail number
origin, dest	Airport codes for origin and destination
air_time	Amount of time spent in the air, in minutes.
distance	Distance flown, in miles.
hour, minute	Time of departure broken in to hour and mins.
time_hour	Timestamp

Exploration ideas:

- Departure delays.
- Best airports in terms of time departure %.
- Aircraft speed analysis.
- On time arrival % analysis.
- Maximum number of flights headed to some particular destination.

Domain: Sports
Project 02: Analyze Football league data

The dataset contains information about Premiere league football from 2012-16.

Variable description

- **FTHG**: home team goals at end of match
- **FTAG**: away team goals at end of match
- **FTR**: match result ([h, a, d] denote [home team victory, away team victory, draw] respectively)
- **HST**: home team shots on target
- **AST**: away team shots on target
- **HC**: home team corner kicks
- **AC**: away team corner kicks
- **HF**: home team fouls
- **AF**: away team fouls
- **HY**: home team yellow cards
- **AY**: away team yellow cards
- **HR**: home team red cards
- **AR**: away team red cards

Exploration ideas

1. Summary Stats: Matches, Teams, Referees, %home win, %away win
2. Relegation Analysis
3. Best/Worst performing teams
4. Playing styles: Fouls, Shots

Domain: Food & Beverages
Project 03: Wine Quality data

The data set is related to different attributes of red wine.

Input variables (based on physicochemical tests)

1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol

Output variable (based on sensory data)

12 - quality (score between 0 and 10)

Exploration ideas

- Data preparation: dividing quality score into 3 different categories, etc.
- Create visualisations to depict how residual sugar, density and alcohol affect the quality of the wine.
- Other variable observations.
- Faulty Wines: Characteristics that can influence wine quality negatively.
- Univariate and bivariate analysis.

Domain: Automobile
Project 04: Automobile data

This dataset contains information about cars

Attribute Information:

Attribute	Attribute Range
1. symboling	-3, -2, -1, 0, 1, 2, 3.
2. normalized-losses	continuous from 65 to 256.
3. make	alfa-romero, audi, bmw, chevrolet,dodge,honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen volvo
4. fuel-type	diesel, gas.
5. aspiration	std, turbo.
6. num-of-doors	four, two.
7. body-style	hardtop, wagon, sedan, hatchback, convertible.
8. drive-wheels	4wd, fwd, rwd.
9. engine-location	front, rear.
10. wheel-base	continuous from 86.6 120.9.
11. length	continuous from 141.1 to 208.1.
12. width	continuous from 60.3 to 72.3.
13. height	continuous from 47.8 to 59.8.
14. curb-weight	continuous from 1488 to 4066.
15. engine-type	dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
16. num-of-cylinders	eight, five, four, six, three, twelve, two.
17. engine-size	continuous from 61 to 326.
18. fuel-system	1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
19. bore	continuous from 2.54 to 3.94.
20. stroke	continuous from 2.07 to 4.17.
21. compression-ratio	continuous from 7 to 23.
22. horsepower	continuous from 48 to 288.
23. peak-rpm	continuous from 4150 to 6600
24. city-mpg	continuous from 13 to 49.
25. highway-mpg	continuous from 16 to 54.
26. price	continuous from 5118 to 45400.

Exploration ideas

- Loading and cleaning data.
- Variable analysis to see its impact on automobile pricing.
- Summary Statistics of different variables.
- Univariate and bivariate analysis
- Make, Curb-weight, Drive wheels analysis.

Domain: Social Network
Project 05: Facebook data

Dataset contains pseudo Facebook data.

Attribute Information:

Userid : ID of user
Age : User's age(years)
dob_day : Day of date of birth(1-31)
dob_year : Year of date of birth
dob_month : Month of date of birth
gender : M/F
tenure : How long have facebook users been on site
friend_count : Total number of friends
friendships_initiated : Friend requests sent
likes : Total number of likes by user
likes_received : Total number of likes received by user
mobile_likes : Number of likes by user(through mobile)
mobile_likes_received : Number of likes recieved by user(through mobile)
www_likes : Number of likes by user(through desktop website)
www_likes_received : Number of likes received by user(through desktop)

Exploration ideas:

- Date of birth analysis
- Friend count analysis
- Tenure analysis
- Data transformations
- Frequency polygons, Boxplots.