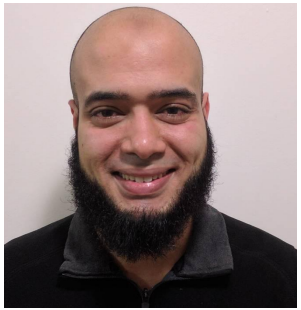
Machine Learning

Exploratory Data Analysis

Mostafa S. Ibrahim
Teaching, Training and Coaching for more than a decade!

Artificial Intelligence & Computer Vision Researcher
PhD from Simon Fraser University - Canada
Bachelor / MSc from Cairo University - Egypt
Ex-(Software Engineer / ICPC World Finalist)



Problem #1: Study House Prices

- Create and explore Kaggle website (ML Competitions)
- Visit the Price Prediction Dataset
- Understand the columns and explore the dataset using the commands we learned
- Study this Exploratory Data Analysis (EDA) on the dataset
  - Stop at : Log Transformation of skewed values
    - And come back later after we finish ‘Modeling Concepts’
  - You will meet new concepts/diagrams/eda. Please study and try!
- **Important tip**
  - The given EDA is an example of a intensive EDA to have in your CV (or a bit less than it)
  - Don’t be the person who just keep generating tables and visualizations
    - Show thinking, insights anc conclusion

Problem #2: Perform Exploratory Data Analysis

- Visit the Superstore Dataset
- Understand the columns and explore the dataset using the commands we learned
  - You will face an error when reading the dataset.
    - Find your way for solving it
    - If failed, see the last page in this document
- Then try to answer the following **questions**
  - If you think a question is not clear, don't ask what the question exactly mean
    - Just do your best exploring around the question goal
  - In practice, data scientists/analysts are supposed to find such basic questions
  - Tip: for each question, think which columns are the relevant ones

Problem #2: Perform Exploratory Data Analysis

- What are the **top selling products** in the superstore?
- What is the sales trend **over time** (monthly, yearly)?
- Which category of products generates the highest revenue and profit?
- Which region generates the most sales?
- What is the impact of discounts and promotions on sales?
- What is the average profit margin for each product category?
    - profit margin = profit/sales (consider it this way)
- Which sub-category of products has the highest demand?

Questions are not mine

Your First Resume Project

- After you get comfortable with running analysis
- I would like you to pick a kaggle project to add in your Resume
  - Title it: Exploratory Data Analysis for <Project-name>
- Use that to present your ability to explore and visualize the dataset and show some of your insights
- Feel free to share with your friends your progress
  - No review from my side. Use people EDAs to get feeling of a good effort

Examples for EDA Datasets

- *In a dataset, search for EDA and sort by most-votes*
- Students Performance in Exams
- Medical Cost Personal Datasets
- COVID-19
- Employee Attrition
- Graduate Admission
- HR Analytics: Job Change
- World University Rankings
- Credit Card Fraud Detection
- Trending YouTube Video Statistics
- Bitcoin Historical Data

"Acquire knowledge and impart it to the people."

"Seek knowledge from the Cradle to the Grave."

The error

- The file has some characters with specific encoding
- Use
- pd.read_csv(Sample - Superstore.csv', **encoding='windows-1252'**)