Final Project Guidelines.

Prepared By: Siman Giri, Instructor: Ronit and Shiv for Herald Center for AI.

Summer, 2025

1 Learning Objectives.

---

1. Explore and preprocess datasets: Clean, transform, and handle missing values and outliers for meaningful analysis.
2. Formulate and test hypotheses: Identify variables, develop null/alternative hypotheses, and apply statistical tests (t-test, ANOVA, chi-square, correlation).
3. Visualize data effectively: Create publication-quality static and interactive visualizations to uncover patterns, trends, and correlations.
4. Build predictive models: Apply regression or classification models, evaluate performance, and interpret results in a business context.
5. Develop interactive dashboards: Present key insights, KPIs, and model results through interactive dashboards using Plotly Dash, Tableau, or Power BI.
6. Document and structure projects professionally: Maintain reproducible code, modular workflows, and clear project documentation.

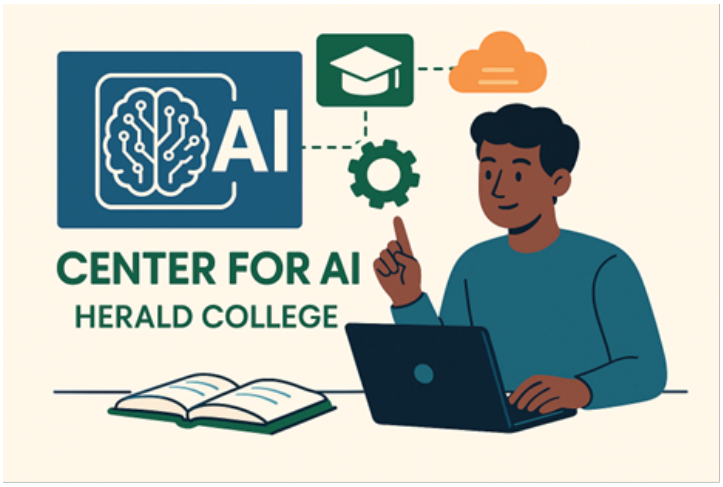

image generated via copilot.

2 Data Analytics Capstone Project.

Goal:

Students will select a real-world dataset related to a business problem, perform exploratory data analysis (EDA) & visualization, build a predictive/statistical model, and present their findings in a dashboard—along with a hypothesis test that validates (or rejects) a key business assumption.

3 Project Workflow:

3.1 Dataset Selection (Business Context Required):

- Students must choose a dataset with a clear business domain (e.g., sales, marketing, finance, supply chain, customer behavior, HR analytics).
- Data can be from Kaggle, government portals, company reports, or self-collected.
- Must have:
  - At least 1 categorical variable and 1 numerical variable.
  - Enough rows (> 500) for modeling and statistical testing.

3.2 Business Problem Definition:

Students must:

- Frame at least one business question that can be answered with data.
- Identify dependent and independent variables.
- State at least one hypothesis (null & alternative).
- **Example:**
  - Business Question: "Does higher marketing spend lead to significantly greater sales?"
  - Hypothesis:
    - $\mathbf{H_0}$: Marketing spend has no significant effect on sales.
    - $\mathbf{H_a}$: Marketing spend has a significant positive effect on sales.

3.3 Data Analysis & Visualization:

1. **Data Cleaning:** Handle missing values, outliers, type conversions.
2. **EDA & Visualization:**
   - Distribution plots, histograms, boxplots.
   - Correlation heatmaps.
   - Trend analysis with time series charts.
   - Comparative plots for categorical variables.
3. All visualizations should be business - friendly, not just technical.
4. Must have at least one interactive plot.

3.4 Dashboard:

Use Power BI, Tableau, or Plotly Dash to:

- Show key metrics (KPIs).
- Create interactive filters (e.g., by product, region, time).
- Highlight trends and anomalies.
- Include at least one visual linked to their hypothesis.

3.5 Hypothesis Testing:

- Use relevant test:
  - t - test/ANOVA $\rightarrow$ comparing means between groups.
  - Chi - square $\rightarrow$ relationship between categorical variables.
  - Correlation significance test.
- Report p - value and confidence intervals.
- Clearly link results to the business question.

3.6 Predictive Modeling:

Students should choose one:

- Linear Regression, Logistic Regression Decision Tree , etc
- Deliverables:
  - Clear explanation of why the model was chosen.
  - Model training and evaluation (R^2, accuracy, RMSE, etc.).
  - Business interpretation of coefficients/results.

4 Appendix 1 - Some Project Ideas:

Following are some of the Suggested research projects student can choose form. Inspired from following student are free to suggest or propose there own project.If the Link for the dataset is broken search for similar dataset.

1. Retail Sales Analytics

1. **Dataset:** Superstore Sales Dataset
2. **Research Questions:**
   - Do certain product categories generate significantly higher profit margins?
   - Is there a difference in average sales between customer segments?
   - Can we predict profit based on discounts, sales, and shipping mode?
   - Do shipping costs vary significantly by region?
3. **Suggested Titles:** Pick One
   - "Uncovering Profit Drivers: A Data-Driven Analysis of Retail Sales Performance"
   - "Sales, Discounts, and Profitability: Insights from Superstore Data"
   - "Do Discounts Pay Off? Statistical and Predictive Analysis of Retail Pricing"

2. HR & Workforce Analytics

1. **Dataset:** IBM HR Analytics Employee Attrition
2. **Research Questions:**
   - Is attrition rate significantly different between job roles or departments?
   - Does monthly income differ significantly between employees who left and those who stayed?
   - Can we predict attrition using satisfaction, income, and tenure?
   - Is overtime more strongly related to attrition than job role?
3. **Suggested Titles:** Pick One.
   - "Why Employees Leave: Predicting Attrition in the Workplace"
   - "From Data to Decisions: Predicting and Preventing Employee Attrition"
   - "The HR Dashboard: Visualizing Attrition Trends and Risk Factors"

3. E-Commerce Operations Analytics

1. **Dataset:** Brazilian E-Commerce by Olist
2. **Research Questions:**
   - Do delivery delays significantly impact customer review scores?
   - Can we predict delivery delays from payment, freight, and product category?
   - Do customers from certain states face longer delivery times?
   - Does payment installment count affect repeat purchase probability?
3. **Suggested Titles:** Pick One
   - "Delivery, Satisfaction, and Sales: Analyzing E-Commerce Operations"
   - "Do Delays Damage Loyalty? A Data-Driven Look at Delivery and Reviews"
   - "Freight, Time, and Reviews: Insights from Brazilian Online Retail"

4. Credit Risk & Finance

1. **Dataset:** Give Me Some Credit Dataset
2. **Research Questions:**
   - Is revolving utilization different for customers with/without delinquency?
   - Can we predict defaults using income, credit history, and utilization rate?
   - Does the number of open credit lines affect default likelihood?
   - Are younger borrowers at higher delinquency risk?
3. **Suggested Titles:** Pick One.
   - "Predicting Credit Risk: A Data-Driven Approach to Customer Evaluation"
   - "From Credit Scores to Defaults: Identifying Financial Risk Factors"
   - "Debt, Income, and Risk: Insights into Loan Delinquency Patterns"

5. Marketing Campaign Analytics

1. **Dataset:** Bank Marketing Dataset
2. **Research Questions:**
   - Is campaign success rate significantly different across job categories?
   - Can we predict term deposit subscription using age, job, and call duration?
   - Does marital status affect subscription probability?
   - Does number of contacts influence success rate?
3. **Suggested Titles:** Pick One.
   - "Who Says Yes? Predicting Customer Responses to Bank Campaigns"
   - "From Cold Calls to Conversions: Analyzing Bank Marketing Effectiveness"
   - "Targeting the Right Customer: Data-Driven Insights for Campaign Strategy"

6. Sports Analytics

1. **Dataset:** FIFA World Cup Matches Dataset
2. **Research Questions:**
   - Do home teams have a significant advantage in win rates?
   - Can we predict match outcomes from past performance?
   - Is there a difference in goals scored between continents?
   - Does player market value correlate with performance?
3. **Suggested Titles:**
   - "Scoring the Odds: Predicting Match Outcomes in International Football"
   - "Home Advantage or Myth? A Statistical Analysis of Tournament Wins"
   - "Player Value vs. Performance: Does Money Buy Goals?"

7. Geopolitics & Development

1. **Dataset:** World Development Indicators (World Bank)
2. **Research Questions:**
   - Is GDP per capita significantly different between OECD and non-OECD?
   - Can we predict life expectancy using GDP, education, and healthcare access?
   - Do higher education expenditures correlate with better HDI scores?
   - Is political stability linked to higher foreign investment?
3. **Suggested Titles:** Pick One.
   - "Development at a Glance: Linking Economic Growth, Education, and Health"
   - "From GDP to HDI: Predicting Human Development from Global Indicators"
   - "Political Stability and Prosperity: A Data-Driven Analysis"

8. Health Economics & Lifestyle

1. **Dataset:** Medical Cost Personal Dataset
2. **Research Questions:**
   - Do smokers have significantly higher medical costs than non-smokers?
   - Can we predict medical charges from age, BMI, and smoking?
   - Does BMI differ significantly by gender?
   - Are costs more related to age or BMI?
3. **Suggested Titles:**
   - "The Price of Health: Predicting Medical Costs from Lifestyle Factors"
   - "BMI, Smoking, and Spending: A Data-Driven Look at Healthcare Costs"
   - "From Habits to Hospital Bills: Statistical Insights into Medical Expenses"

5 Appendix - 2 - Sample Project.

Following presents the sample project built using Titanic Dataset for demonstration purpose only.

5.1 Suggested Project Structure:

Highly Recommended:

```
student-project/
+-- data/                      # Raw and processed datasets.
|   +-- raw/
|   +-- processed/
|
+-- notebooks/          # Jupyter notebooks for EDA & Prototyping
|   +-- 01_data_exploration.ipynb
|   +-- 02_hypothesis_testing.ipynb
|   +-- 03_modeling.ipynb
|   +-- 04_dashboard_prototype.ipynb
|
+-- src/              # Python scripts (modularized codebase)
|   +-- data_preprocessing.py
|   +-- eda.py
|   +-- hypothesis.py
|   +-- modeling.py
|   +-- visualization.py
|
+-- dashboard/     # Dash app for final dashboard
|   +-- app.py
|   +-- layouts.py
|   +-- callbacks.py
|   +-- assets/
|
+-- reports/     # Results and documentaions.
|   +-- figures/ # saved plots, charts
|   +-- final_report.pdf # Optional (Not Required)
|
+-- tests/  # (Optional) Any tests you may have performed.
+-- requirements.txt # Project Dependencies.
+-- README.md # Project Overview
+-- .gitignore
```

Less Recommended But Accepted:

```
student-project/
+-- data/                      # Raw and processed datasets.
|   +-- titanic.csv
|
+-- app/              # Python scripts (modularized codebase)
|   +-- __init__.py
|   +-- data_processing.py
|   +-- layout.py
|   +-- callbacks.py
|   +-- visualization.py
+-- main.py
+-- requirements.txt # Project Dependencies.
+-- README.md # Project Overview
+-- .gitignore
```

The provided sample project follows above structure (I am not happy about it, but it works).

5.2 Setup Instructions

1. Clone the repository

```
git clone https://github.com/girisiman/davcoursedemoproject.git
cd <your-repo-name>
```

2. Create and activate a virtual environment

Using Python venv:

```
python -m venv venv
# Windows
venv\Scripts\activate
# macOS/Linux
source venv/bin/activate
```

Using Conda:

```
conda create -n titanic_env python=3.10
conda activate titanic_env
```

3. Install dependencies

```
pip install -r requirements.txt
```

4. Run the dashboard

```
python main.py
```

Open your browser and go to: `http://127.0.0.1:8050` to view the dashboard.

Notes

- This is a simple demo to illustrate the minimum requirements for the final project.
- Students can replace `titanic.csv` with their own dataset for practice.
- For more details about the course, visit: Data Analytics & Visualization Course.

The - End -