DS 204 — Analysis for Data Science

Final Project

Project Overview

Each group will select a real-world dataset (from sources such as Kaggle, UCI, Data.gov, etc.), produce a detailed **data quality report**, identify **2–3 meaningful associations** between features using multiple correlation coefficients (Pearson, Spearman, Kendall), fix data quality issues in code, and propose a prediction approach (focusing on linear regression options where appropriate).

You will deliver a reproducible Jupyter Notebook, a compact written analysis (PDF), a fixed CSV dataset, presentation slides, and a recorded group presentation.

Learning Outcomes

After completing the project students will be able to:

- Produce a professional data-quality profile and interpret it.
- Select and defend appropriate correlation measures (Pearson, Spearman, Kendall).
- Identify and explain associations between features and decide which is explanatory vs response.
- Detect and fix common data quality problems programmatically.
- Recommend and justify a regression approach (or alternative) and evaluate it.
- Present results clearly in writing and in a recorded presentation.

Step-by-step Instructions:

1. Search and import a dataset from authentic online sources like Kaggle, UCI Machine Learning Repository etc. Make sure the dataset has enough numeric features to be analyzed. Otherwise, you will have to encode categorical features to numbers.
2. Produce a data quality report along with multiple types of correlation coefficients calculated (Pearson, Kendall's, Spearman's). You may use ydata_profiling package or DataPrep whichever you are comfortable with.

3. Use the Data Quality Report to find out associations among features. Find at least 2 – 3 associations (2 – 3 pairs of features which you feel are correlated with each other).
4. Create a PDF document to report your findings. In your Analysis report:
    a. Explain each association along with an indication of which feature can be the explanatory variable and which one can be a possible response variable.
    b. Explain and interpret the correlation coefficient value. What is the value? How strong is the correlation? What is your interpretation of the correlation coefficient? Also, explain which correlation coefficient was used to analyze the association and why?
    c. Once you confirm that there is a probable association between the noted features, analyze and explain what sort of data quality issues (if any) are present within the features. They can range from missing values, duplication in the dataset, skewness/outliers, etc. Provide necessary proof to support your findings.
    d. If there are any data quality issues present, explain your approach to fix them along with a solid rationale behind your choice of fix.
    e. Suggest what sort of prediction models can be created using the associations you found out and how they can be used in real life scenarios. Specifically focus on linear regression-based models.
5. Go back to your Python notebook produced in Section 2. Add necessary code to fix the data quality issues identified and apply suggested approaches from Section 4.
6. Export your fixed dataset file to a .csv file. Name it **"Dataset-Fixed.csv".**
7. Based on your suggestions in section 4(e), choose one of the best possible prediction models which can be implemented through Linear Regression. Provide rationale behind your choice (for example, a scatter plot showing linear relationship between variables of choice). Add this information to your analysis report produced in Section 4. If a linear relationship is not present between any of the variables for your chosen associations, then suggest a best possible prediction model for non linear relationships and provide rationale.
8. Create a presentation (using any tool of choice like MS PPT, Canva, etc.) and present your work in a video recorded presentation with all the group members present in the video. Make sure your cameras are switched on during the presentation. Record a **15-20 minute presentation** explaining the following:
    a. Introduce your dataset
    b. Discuss the associations that you identified along with the interpretation of correlation coefficient and identification of explanatory and response features.

c. Discuss any data quality issues present in the features present within your associations. Discuss the rationale behind your findings.
d. Discuss approaches to fix data quality issues that were identified. Discuss rationale behind your choice of fixes.
e. Discuss the potential options for ML Models that you found out and suggest the best possible option for Linear Regression Implementation.
f. In addition, please make sure to show outputs and necessary code within your presentation.

The presentation can be recorded using any online meeting tool (for example, Zoom). Upload your video recording on Google Drive and share the link to access the video as a private comment along with your project submission files on Google Classroom. Also, put the same video access link at the end of your analysis report document. **Make sure to change the sharing settings of your video to "Anyone with the link can access" so the instructor can access it.**

Detailed Guide on Project Instructions:

Dataset Selection (Requirements)

- Select a dataset from a public, authentic source (Kaggle, UCI, Data.gov, Google Dataset Search, etc.).
- The dataset **must contain enough numeric features** for correlation/linear analysis. If you have categorical variables you wish to use, you must encode them to numeric with proper justification (one-hot, label encoding, target encoding, etc.).
- Document the dataset's provenance (download link), domain, size (rows × columns), and why it is suitable for analysis.

Required Analyses & Artifacts (Deliverables)

Submit the files mentioned below and upload them through Google Classroom in the submission link for Final Project. Place the recorded presentation link in the PDF analysis and as a private comment on Google Classroom.

1. **Jupyter Notebook** (<GroupNumber>_DS204_Notebook.ipynb) — fully reproducible, all code executed, with markdown explanations. Notebook must include:
    - Dataset description and metadata.

- Data quality profiling (ydata_profiling or DataPrep) with calculation of **Pearson, Spearman, and Kendall** correlation coefficients.
- Feature selection and reasons.
- Visual proofs (scatter plots, rank plots, residuals, etc.) to support choice of associations.
- All data cleaning/fixing steps (code + rationale).
- Modeling suggestion (linear regression or justified alternative).

2. **Data quality report** profile.html (generated by ydata_profiling or DataPrep). Submit the HTML file.
3. **Fixed dataset CSV** Dataset-Fixed.csv — the cleaned dataset exported after fixing the data quality issues.
4. **Analysis report (PDF)** <GroupNumber>_DS204_Analysis.pdf — a concise written report (6–10 pages recommended) containing:
   - Dataset source & description.
   - Data quality findings and evidence.
   - The 2–3 associations (each: coefficient values, plots (where applicable), explanation of explanatory vs response variable).
   - Data cleaning steps taken and rationale.
   - Chosen prediction model suggestion and why (with short justification).
   - Link to the recorded presentation (must be at the end of the PDF).
5. **Presentation slides** <GroupNumber>_DS204_Slides.pptx (or PDF).
6. **Recorded group presentation** (15–20 minutes). Cameras must be on and all group members should appear. Upload to Google Drive (or another provider) and **set sharing such that the instructor can access via link.** Put the link in the PDF and in the Google Classroom submission.

Notebook / Code Expectations (Format & Reproducibility)

- Use cells with explanatory markdown: each major step must have a short description and expected outcome.

- Add a header cell listing **package versions** used (pandas, numpy, scipy, sklearn, statsmodels, ydata_profiling / dataprep) so results can be reproduced.
- Avoid hard-coded paths. Use relative paths or a single data/ folder (e.g., data/raw.csv, data/Dataset-Fixed.csv).

Data Quality Report — What to Show

Your profile (ydata_profiling or DataPrep) must include and the notebook must discuss:

- Missing value counts and percentages per column.
- Duplicates.
- Data types and unique value counts.
- Distribution summaries (mean, median, std, skewness).
- Outlier detection summary.
- Pairwise correlations summary (as the profiling tool shows) – including Pearson, Spearman and Kendall's correlation coefficients.

These outputs should be used to pick candidate feature pairs for association analysis.

Association Analysis — Requirements & How to Present Results

For **2–3 feature pairs** you must:

1. State which pair you are testing and why (domain logic).
2. Analyze **Pearson**, **Spearman**, **Kendall** coefficients and report which coefficient is most appropriate and why.
3. Show a scatter plot where appropriate.
4. Identify which feature you treat as **explanatory (X)** and which as **response (Y)** and justify.
5. If you propose a linear regression model explain the rationale. If assumptions fail, explain and propose an alternative model. (for example, polynomial regression, logistic regression etc.)

Quick Interpretation Guide (Use in Report)

- Strength (heuristic) of correlation based on coefficient values:
    - 0.00–0.19: Very weak
    - 0.20–0.39: Weak
    - 0.40–0.59: Moderate
    - 0.60–0.79: Strong
    - 0.80–1.00: Very strong

Which Correlation to Use — Practical Guidance

- **Pearson**: use when both variables are continuous and have a *linear* relationship and no extreme non-normality. Reports linear association strength.
- **Spearman**: rank-based, use when variables are not normally distributed or the relationship is monotonic but not linear, or for ordinal data.
- **Kendall**: also rank-based; more robust on small samples and with many tied ranks.

Data Cleaning (What to Include & Recommended Methods)

For each data quality issue you identify, you must:

1. **Show evidence** in the notebook (code output / plots).
2. **Explain your chosen approach** and rationale.

Common issues & recommended treatments (show before/after and % affected):

- **Missing values**
    - Report count & % per column.
    - Small %: consider mean/median (numeric) or mode (categorical).
    - Large %: consider dropping the column with justification.
- **Duplicates**
    - Show duplicate rows count; drop duplicates where appropriate.

- **Outliers / skewness**
    - Use boxplots / z-score / IQR method; options: winsorize, transform (log/Box-Cox), or remove (with justification).
- **Incorrect data types**
    - Convert strings to datetimes, coerce numeric fields; show checks before/after.
- **Feature engineering**
    - Create derived features (ratios, logs, bins) where justified by domain logic.

Always keep raw data unchanged and save cleaned CSV as **Dataset-Fixed.csv.**

Modeling Suggestions / Prediction (Brief)

- If linear assumptions are reasonably satisfied:
    - Suggest a simple linear regression model based on your associations.
- If linear assumptions fail:
    - Try suggesting transformations (log, Box-Cox), polynomial features, or tree-based models (Decision Tree, Random Forest). Provide quick comparative metrics and justify the choice.
- Focus on **interpretability** and reproducibility; you do not need to build an extensive production pipeline.

Presentation (Recorded) & Submission Checklist

Recorded Presentation (15–20 min)

Structure:

1. Dataset intro & provenance (1–2 min).
2. 2–3 association findings (plots + interpretation) (5–7 min).
3. Data quality issues & fixes (3–4 min).
4. Modeling suggestion & rationale (3–4 min).

5. Conclusion (1–2 min).

Requirements:

- All team members must appear on camera.
- Slides should be clear and minimal text; figures must be readable.
- Upload video (Google Drive or any other platform) and include accessible link in the PDF report and on Google Classroom submission comments.

Submission checklist (what to upload)

1. <GroupNumber>_DS204_Notebook.ipynb (executable).
2. profile.html (data quality profile).
3. Dataset-Fixed.csv (cleaned dataset).
4. <GroupNumber>_DS204_Analysis.pdf (analysis report with video link for presentation).
5. <GroupNumber>_DS204_PresentationSlides.pptx (or PDF).
6. Link to recorded presentation as a private comment

Submit everything on Google Classroom under project submission link.

Grading Rubric

- Dataset choice & documentation — 10%
- Data quality report & diagnosis — 20%
- Data cleaning & fixed dataset — 20%
- Association analysis & interpretation — 25%
- Modeling suggestion & evaluation — 10%
- Presentation & professionalism — 10%
- Reproducibility & notebook quality — 5%

Tips & Common Pitfalls

- Do **not** cherry-pick results to make correlations look stronger; show checks and caveats.
- Keep raw and cleaned copies — never overwrite raw data.
- Write short clear captions for every plot; the grader must understand a plot without reading code.
- Quantify changes when imputing or removing data (e.g., “5% of rows had missing y and were dropped”).
- Put long code in the notebook; keep PDF focused on interpretation and results.
