DATA 4381: Final Report Fall 2024
Project Title: Restaurant Health Inspection Score Prediction
Prepared by: Thao Wells
Date: 12/07/2024

Introduction
Over the course of this semester, I have worked on a project aimed at predicting health inspection scores for restaurants in San Francisco. The project entailed exploring and preprocessing the dataset, conducting exploratory data analysis (EDA), applying Principal Component Analysis (PCA), and building predictive models using machine learning techniques. The journey involved overcoming technical challenges, interpreting insights from data, and continuously refining the methodologies to enhance model performance.

---

1. Clarity and Structure
This report synthesizes the progress made throughout the semester on the restaurant health inspection score prediction project. Information is structured to outline milestones achieved, challenges faced, and plans for further development.

---

2. Progress and Milestones

Tasks Completed:
1. Data Preprocessing:
   - Cleaned and formatted the dataset, dropping irrelevant columns (e.g., `business_name`, `business_address`, `business_location`).
   - Applied One-Hot Encoding to categorical variables.
   - Addressed missing values by balancing data retention and cleaning integrity.

2. Exploratory Data Analysis (EDA):
   - Conducted EDA to uncover relationships between features and inspection scores.
   - Created visualizations such as bar charts, pie charts, and choropleth maps to illustrate data distribution and geographic patterns.

3. Principal Component Analysis (PCA):
   - Reduced dimensionality and identified components driving variance.
   - Explored geographic and location-based influences on inspection scores.

4. Model Development and Evaluation:
   - Implemented models: Random Forest Classifier (RFC) and Support Vector Machine (SVM).

- Evaluated model performance using metrics such as accuracy, precision, recall, F1 score, and confusion matrices.

Key Metrics and Results:
Random Forest Classifier:
- **Precision:** 0.737
- **Recall:** 0.730
- **Accuracy:** 0.730
- **F1 Score:** 0.726

Support Vector Machine (SVM):
- Challenges in execution due to memory issues; results are pending optimization.

PCA Component Loadings:

Variable	PC1	PC2	PC3
Neighborhoods	0.602743	0.205155	0.219469
SF Find Neighborhoods	0.602743	0.205155	0.219469
Business Latitude	-0.383798	0.297845	0.270317
Business Longitude	0.014583	0.654894	-0.083413
Current Police Districts	0.234036	-0.594277	0.060511
Inspection Score	0.042912	-0.178994	0.242202
Current Supervisor Districts	-0.250735	0.002090	0.606602

Key Insights:
- **PC1:** Location features (Neighborhoods, SF Find Neighborhoods) are significant drivers of variance.
- **PC2:** Geographic coordinates (business_longitude) and police districts influence inspection outcomes.
- **PC3:** Supervisor districts and inspection scores correlate with variance.

---
3. Problem-Solving and Challenges

Challenges Encountered:
1. **Data Preprocessing:**
   - Balancing missing value handling with data retention.

2. **PCA Implementation:**
   - Interpreting principal component loadings required extensive research and refinement.

3. **Model Development:**
   - SVM implementation faced memory constraints; future work includes optimizing feature scaling.

4. **Visualization:**
   - Choropleth map misalignments resolved by confirming data-layer compatibility.

These challenges enhanced my technical skills and deepened my understanding of data science methodologies.

---

4. Technical Depth and Accuracy
The project involved the following technical components:
- **Data Preprocessing:** Ensured robust data preparation using Pandas and NumPy.
- **Visualization:** Created visualizations using Matplotlib to effectively communicate data insights.
- **Machine Learning:** Applied Scikit-learn to implement models and evaluate performance.
- **PCA:** Used PCA to reduce dimensionality and extract meaningful patterns from data.

---

5. Future Plans and Goals

Tasks for Future:
1. **Model Optimization:**
   - Refine preprocessing and feature scaling to improve SVM performance.
   - Explore other algorithms, such as XGBoost and Gradient Boosting, to enhance predictive power.
   - Investigate the confusion matrix for Random Forest Classifier to identify misclassified classes.

2. **Feature Engineering:**
   - Integrate external datasets (e.g., crime rates, highway proximity) to explore correlations.
   - Refine feature selection to reduce noise and improve model accuracy.

3. **Visualization Enhancement:**
   - Develop interactive dashboards for better presentation of insights.

4. **Deployment Preparation:**
   - Prepare the model for real-world application by developing an API or web interface.

Conclusion
The Restaurant Health Inspection Score Prediction project has provided invaluable learning experiences. By combining data preprocessing, EDA, PCA, and machine learning models, I

have laid the foundation for predicting inspection scores. While challenges remain, the progress made offers a strong basis for future enhancements and applications.