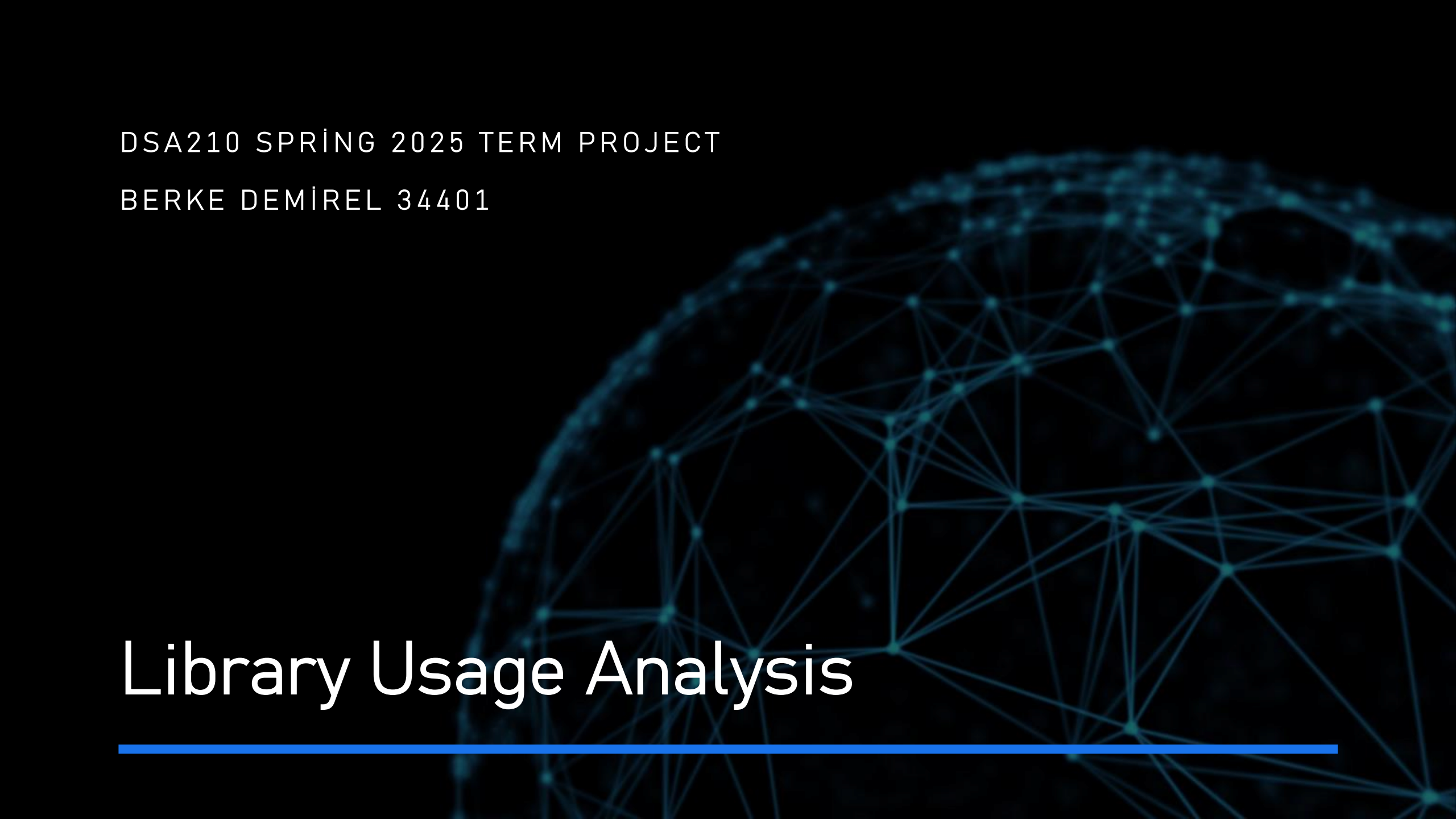
DSA210 SPRİNG 2025 TERM PROJECT

BERKE DEMİREL 34401



Library Usage Analysis

Motivation

Public libraries serve as critical community resources, offering free access to education, technology, and cultural programs. However, library engagement—such as book checkouts or event attendance—varies widely across neighborhoods. These disparities often reflect underlying socioeconomic factors like income levels, education, and population density. By analyzing library usage alongside demographic data, we can uncover patterns that guide fair resource distribution and advocate for data-driven policy decisions.

This project focuses on answering two key questions:

1. How do community demographics correlate with library engagement metrics (e.g., checkouts, program attendance)?
2. Can we predict library usage patterns using socioeconomic indicators?

Data Sources & Preprocessing

1. IMLS Public Library Survey (2022)

Key Features:

- TOTCIR: Total circulation of materials
- VISITS: Total annual library visits
- REGBOR: Number of registered users
- TOTSTAFF: Total paid FTE employees
- GPTERMS: Internet computers used by general public
- HRS_OPEN: Total annual public service hours for all service outlets
- TOTPRO: Total number of synchronous program sessions
- TOTATTEN: Total attendance at synchronous programs
- TOTINCM: Total operating revenue
- TOTOPEXP: Total operating expenditures
- POPU_UND: Unduplicated population of the legal service area for the library
- ZIP_CODE: Administrative ZIP code of the library system

2. US Census ACS 5-Year Estimates (2022)

Key Features:

- MEDIAN_INCOME: Median household income by ZIP Code Tabulation Area (ZCTA)
- BACHELORS_PERCENT: Percentage of population with a bachelor's degree or higher
- UNEMPLOYMENT_RATE: Unemployment rate
- ZIP_CODE: ZIP Code Tabulation Area (ZCTA)

Data Sources & Preprocessing

-Datasets were merged on ZIP_CODE

-Some entries in the data that had the values -9, -3, -1, 0 which indicated suppressed/unusable data were converted to NaN

-Rows with NaN population were dropped from the dataset

-Binary "_IMPUTED" indicator columns were created for each feature, 1 indicates the value is imputed, 0 indicates otherwise

-The columns were imputed using the median since some of the features were skewed

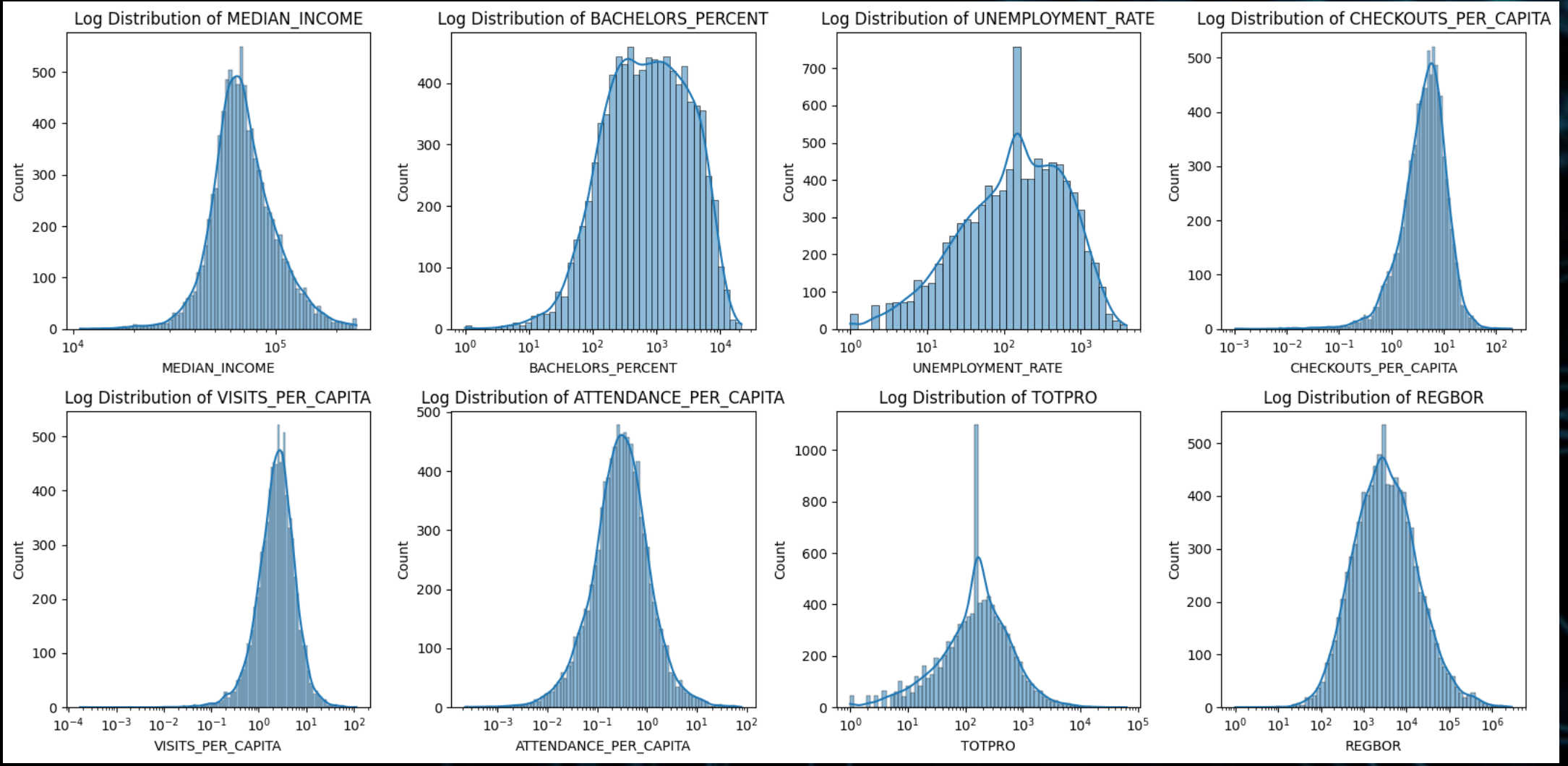
-Three new features were derived using the existing features, namely:

"CHECKOUTS_PER_CAPITA", "VISITS_PER_CAPITA", "ATTENDANCE_PER_CAPITA" from the original features "TOTCIR", "VISITS", "TOTATTEN" by dividing them by the respective population values

-Those three original features were then dropped from the dataset

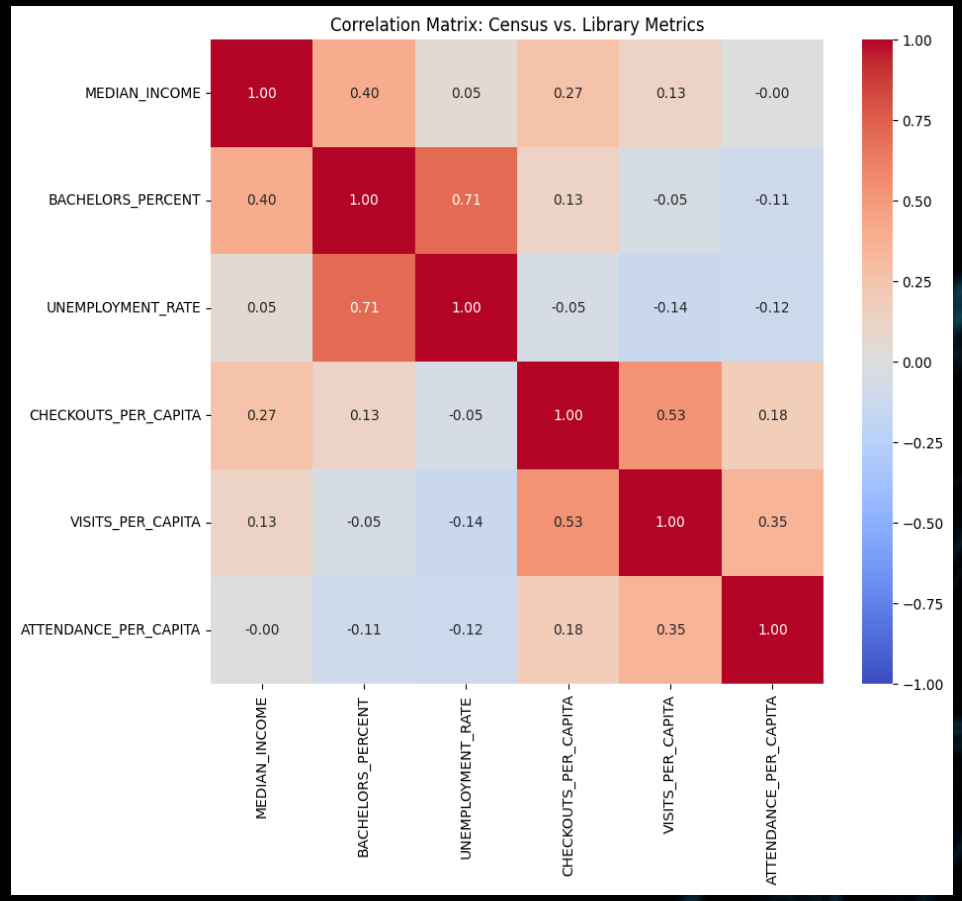
Exploratory Data Analysis

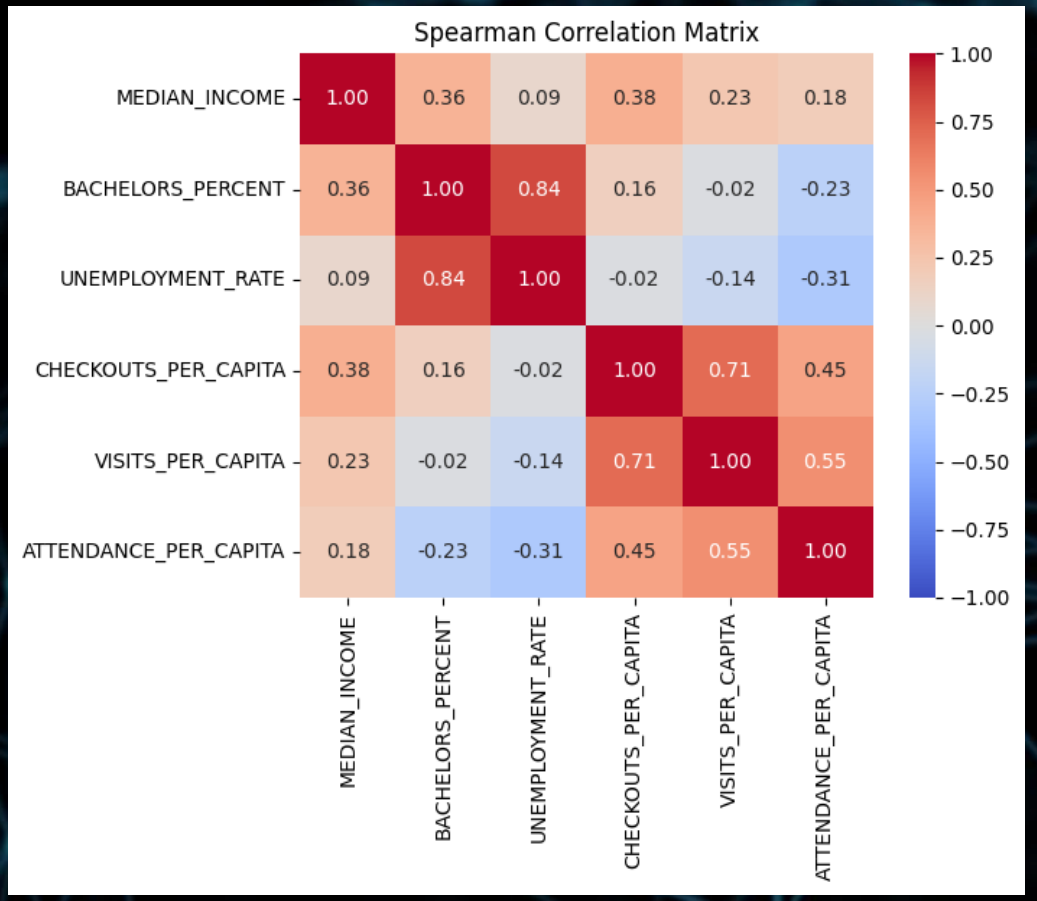
Log distributions of some of the features:



Exploratory Data Analysis

Correlation matrices: Pearson & Spearman Correlations

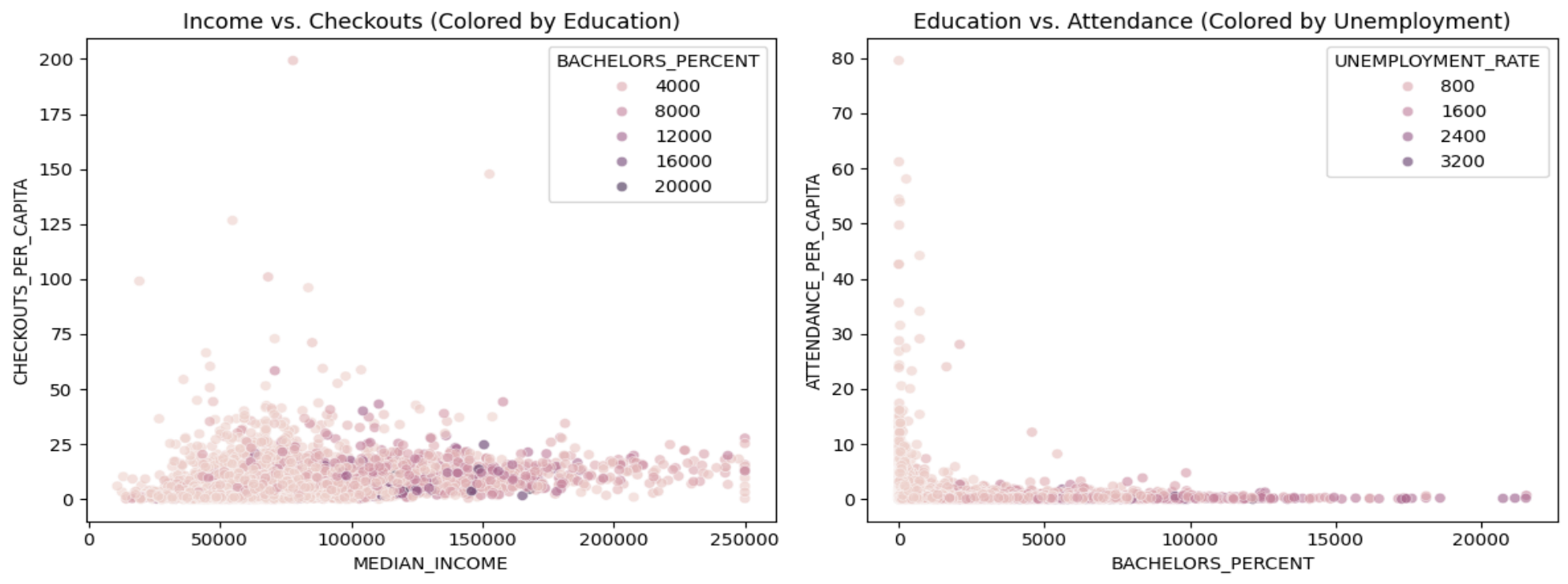




Strong correlations: UNEMPLOYMENT_RATE & BACHELORS_PERCENT, VISITS_PER_CAPITA & CHECKOUTS_PER_CAPITA

Exploratory Data Analysis

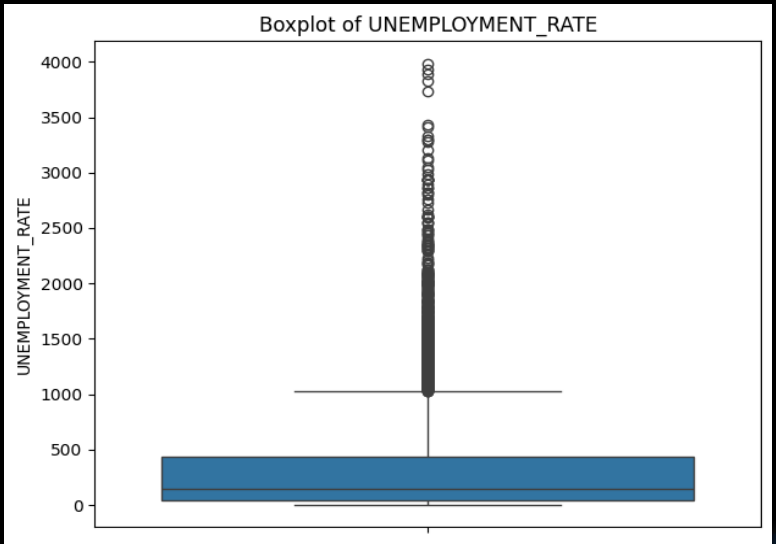
Scatter plots: Income vs. Checkouts & Education vs. Attendance

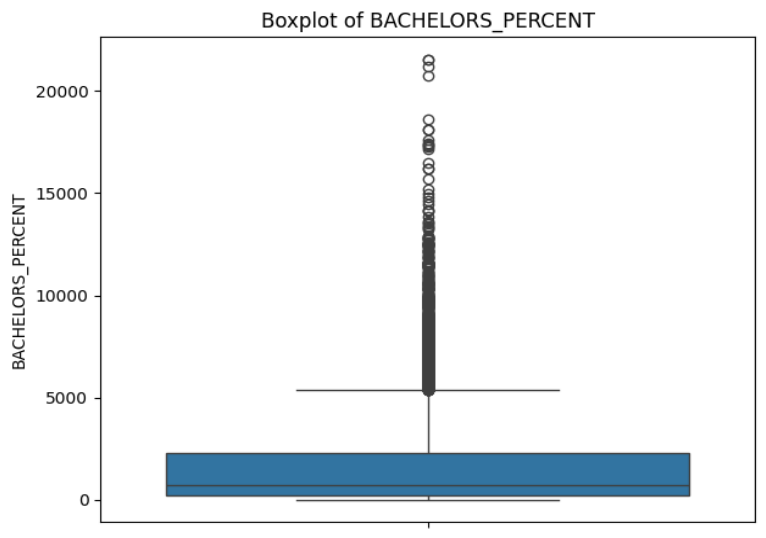


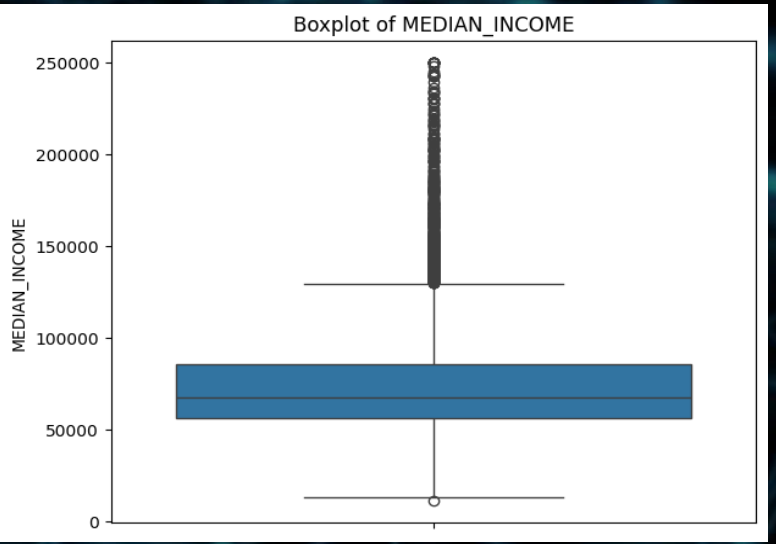
By the first scatter plot, it can be concluded that there is some correlation between median income versus checkouts per capita, although it is not strong. By the second scatter plot, it can be concluded that there is a negative correlation between education levels and program attendance per capita, although just like the first plot, this relationship is not strong.

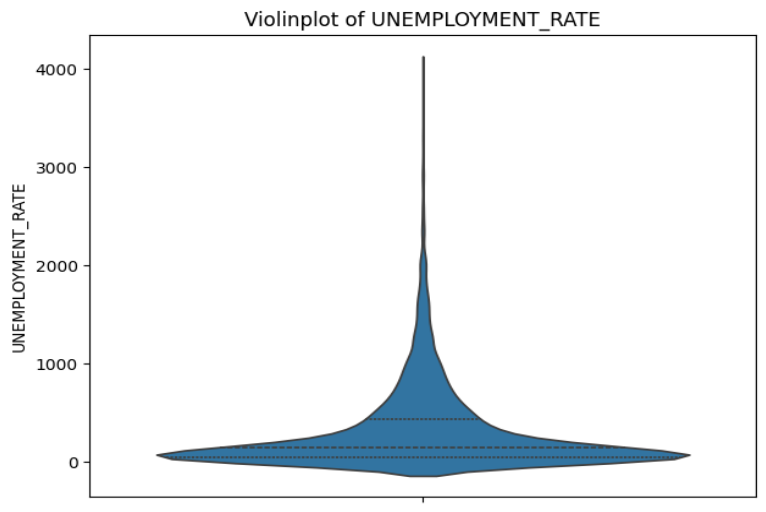
Exploratory Data Analysis

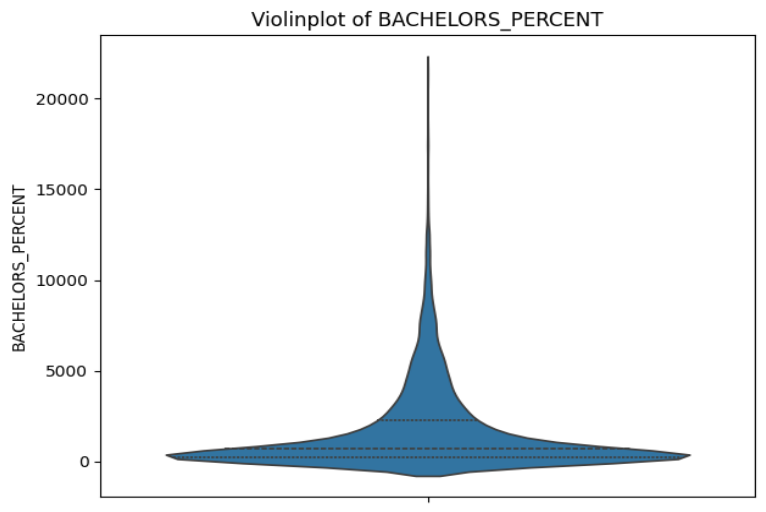
Boxplot and Violinplots of UNEMPLOYMENT_RATE, BACHELORS_PERCENT & MEDIAN_INCOME

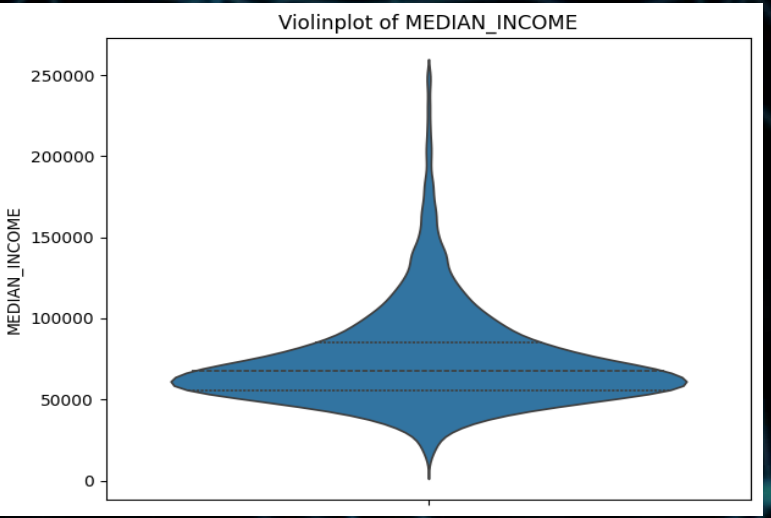










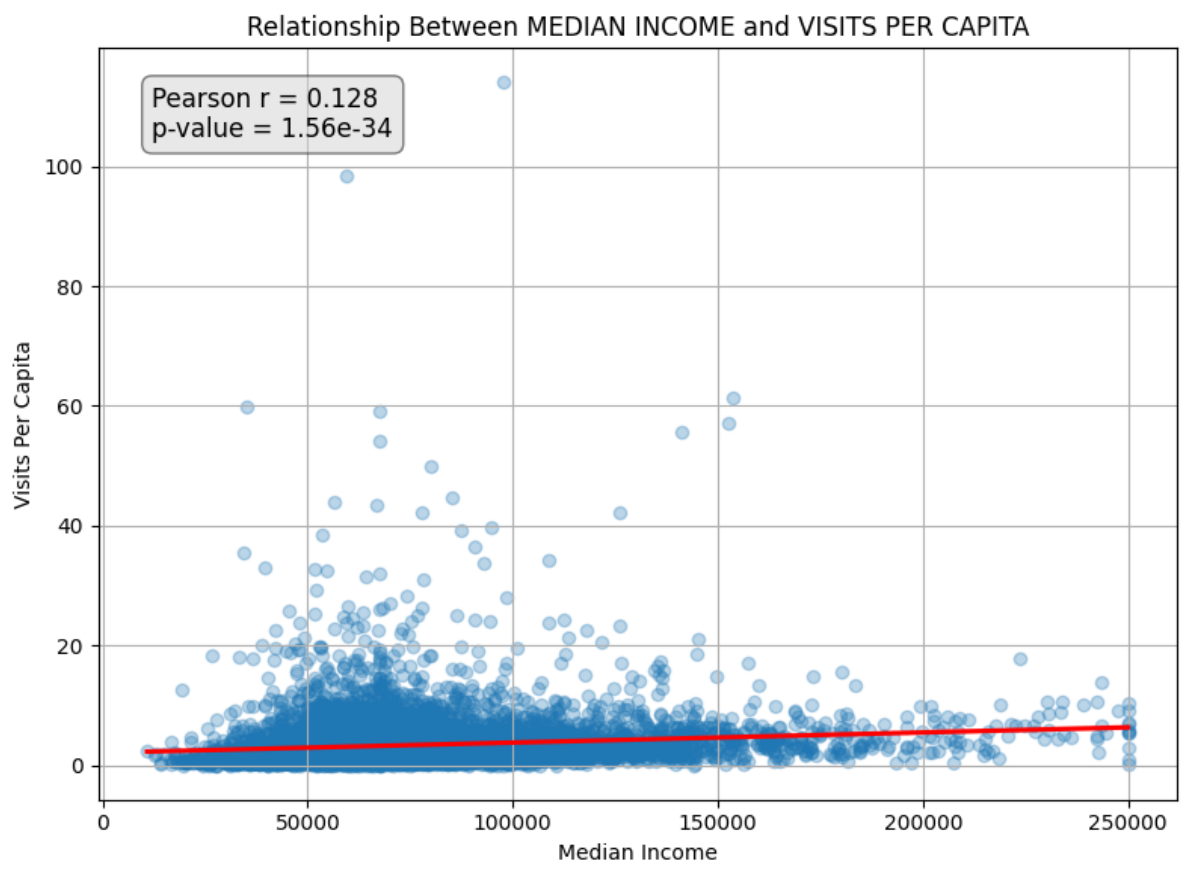


Hypothesis Tests

Hypothesis Test 1: Median Income and Library Visits

-H_0: Median household income and visits per capita are uncorrelated.

-H_1: They are positively correlated.



Since p-value ≪ 0.05, we reject the null hypothesis: There is statistical evidence of a non-zero association, but since r= 0.128, this result is not practically meaningful, i.e. income only very slightly predicts library visits per capita.

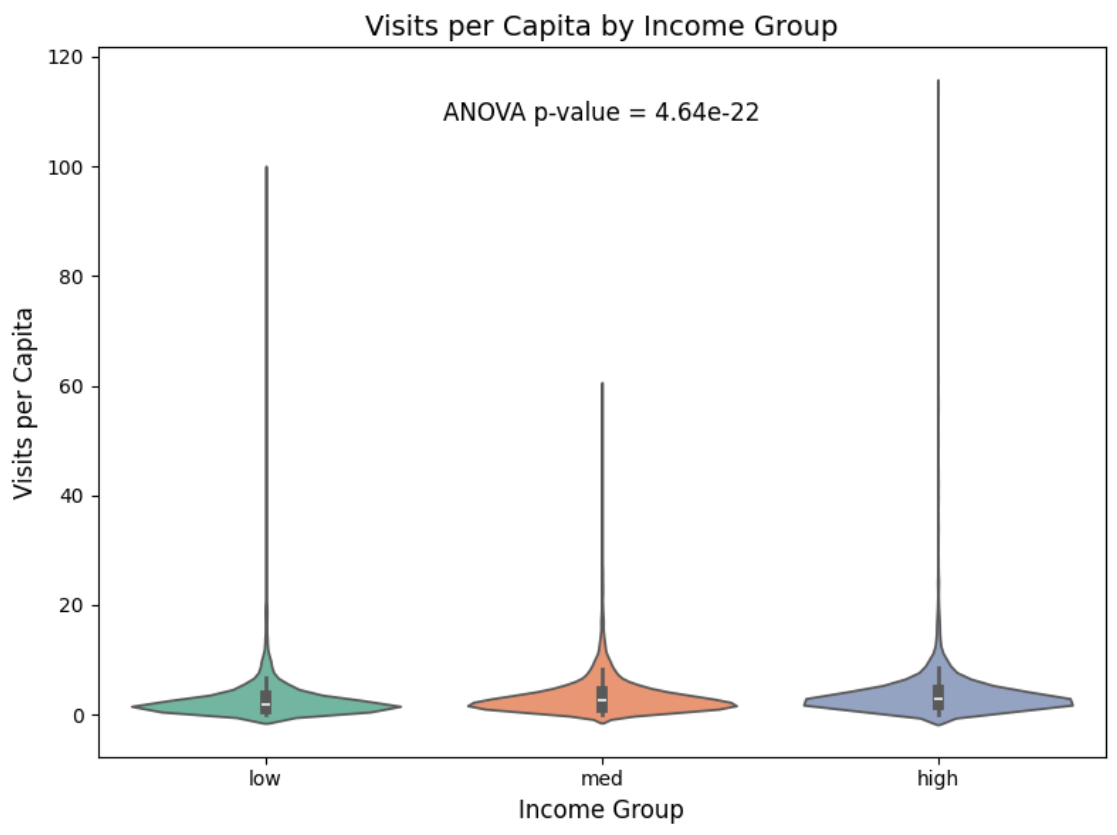
Hypothesis Tests

Hypothesis Test 2: One-Way ANOVA

Question: Do low / medium / high income ZIPs differ in mean visits per capita?

-H_0: All three group means of VISITS_PER_CAPITA are equal

-H_1: At least one group mean differs



Since p-value < 0.05, we reject the null hypothesis. There is a statistically significant difference in VISITS_PER_CAPITA across income groups (low/med/high).

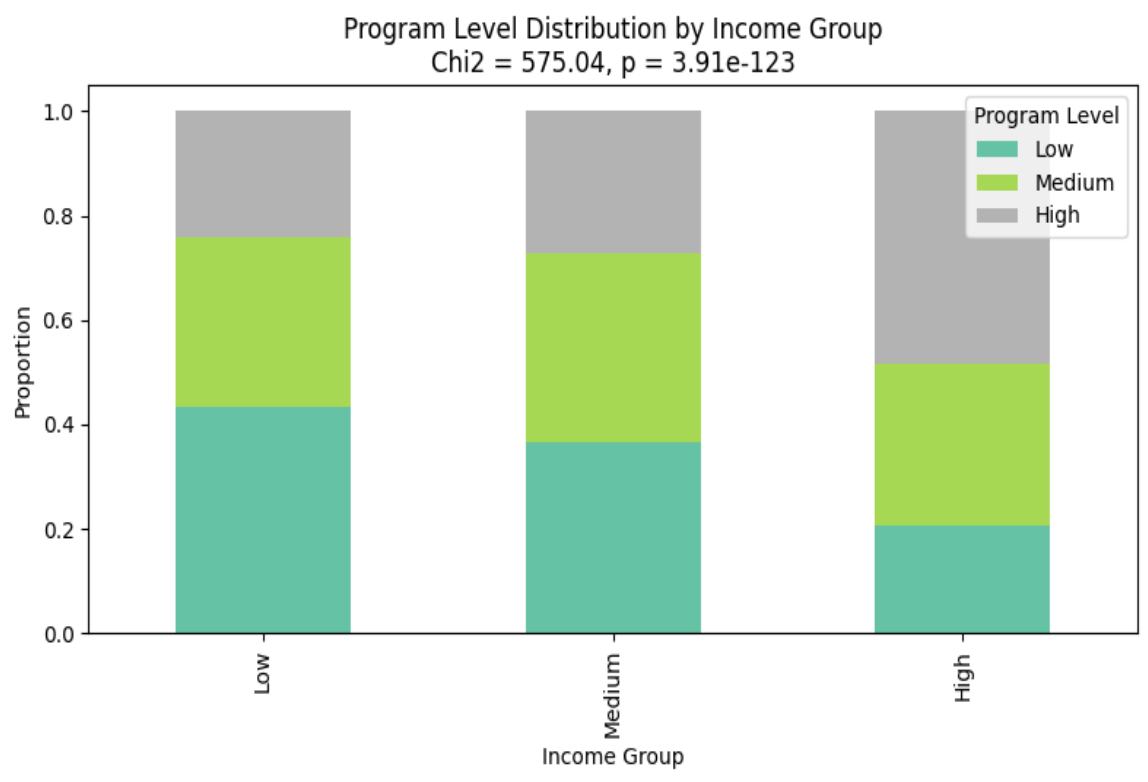
Hypothesis Tests

Hypothesis Test 3: Chi-Square Test of Independence

Question: Is there an association between ZIP code income level (Low / Medium / High) and library program offerings?

-H_0: Income group and program offering level are independent.

-H_1: Income group and program offering level are associated.



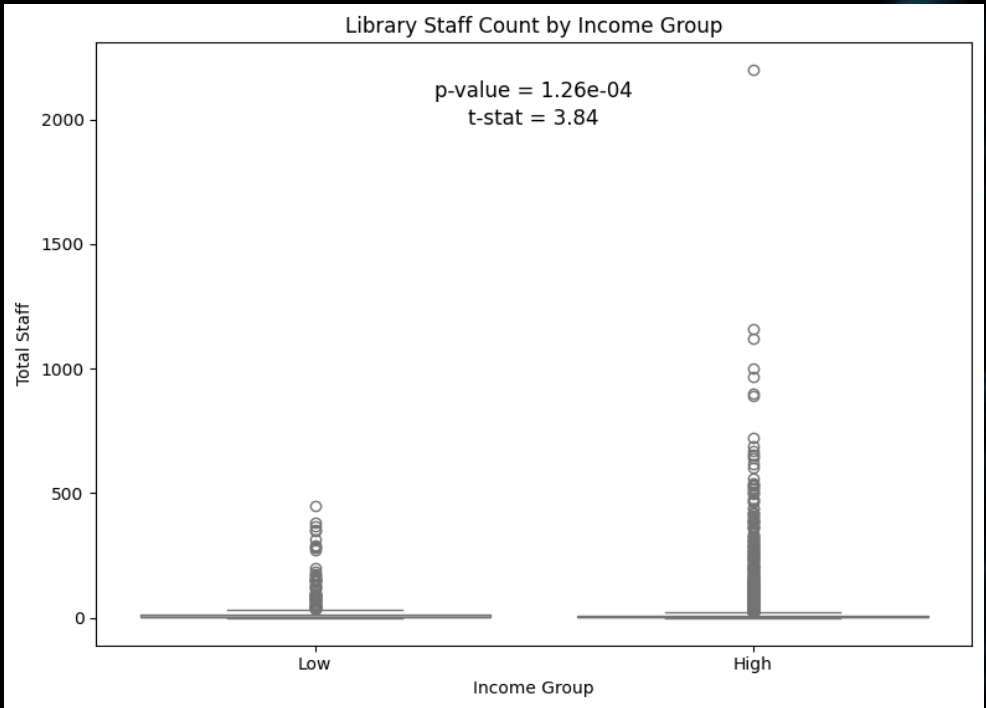
Since p_value ≪ 0.05, and even ≪ 0.001, we reject the null hypothesis. Library program levels tend to differ substantially depending on whether a ZIP code is low-, medium-, or high-income.

Hypothesis Tests

Hypothesis Test 4: Two-Sample T-Test

Question: Do libraries in high-income ZIP codes have significantly more library staff (TOTSTAFF) than those in low-income ZIP codes?

- H_0: There is no significant difference in the mean number of library staff between high-income and low-income ZIP codes.
- H_1: There is a significant difference in the mean number of library staff between high-income and low-income ZIP codes.



Since p_value ≪ 0.05, we reject the null hypothesis. The difference in the mean number of library staff between two income groups is significant, although looking at the box plot comparison, this result is not practically meaningful.

Handling Collinearity

Used Variance Inflation Factor (VIF) with threshold 6 to exclude the features that exhibited high mutlicollinearity before training the machine learning models.

```
=== VIF before dropping features ===
```

	feature	VIF
9	TOTOPEXP	112.845552
8	TOTINCM	108.191393
5	TOTSTAFF	21.228151
10	POPU_UND	9.054362
6	HRS_OPEN	6.054702
3	REGBOR	5.438960
4	GPTERMS	4.501981
7	TOTPRO	3.376735
2	BACHELORS_PERCENT	2.901187
0	UNEMPLOYMENT_RATE	2.374683
1	MEDIAN_INCOME	1.411176
11	VISITS_PER_CAPITA	1.188871
12	ATTENDANCE_PER_CAPITA	1.158057

```
Dropping:  ['TOTOPEXP', 'TOTINCM', 'TOTSTAFF', 'POPU_UND', 'HRS_OPEN']
```

5-Fold CV for Linear Regression

Target variable to predict was CHECKOUTS_PER_CAPITA. Numeric features were scaled with StandardScaler(). Used k-Fold cross validation (k = 5) to train and test the Linear Regression model.

Results:

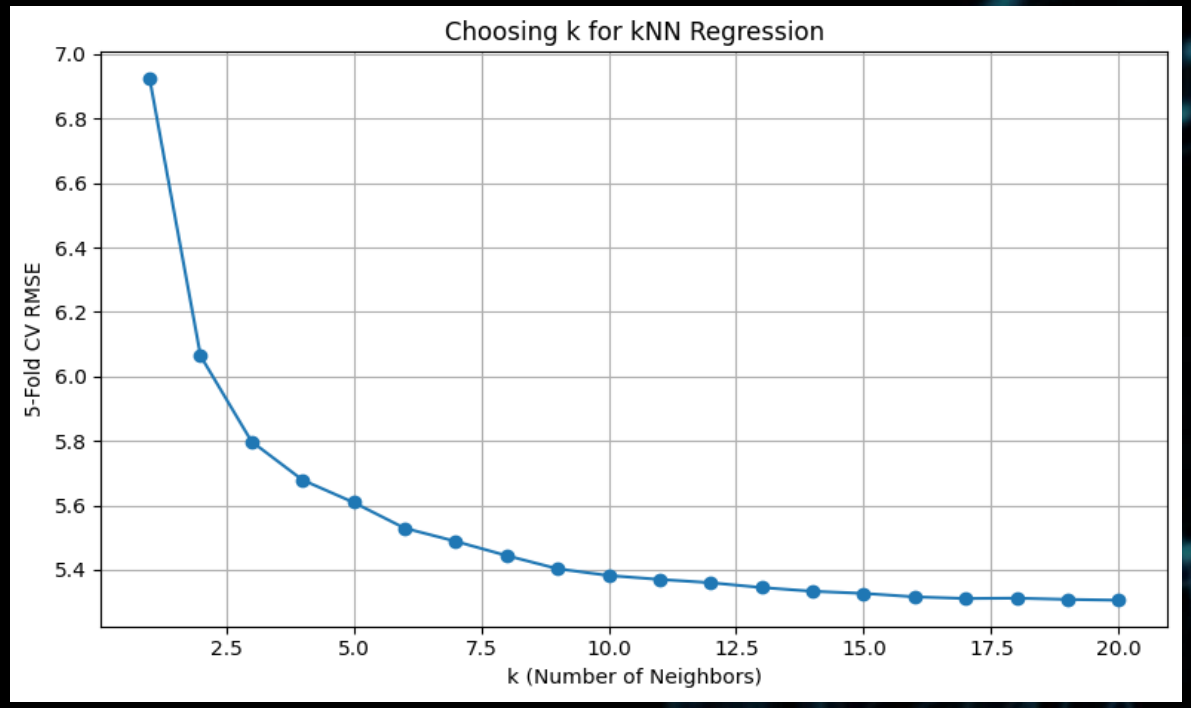
```
Fold 1 | Linear Regression -> RMSE 5.070, R² 0.299
Fold 2 | Linear Regression -> RMSE 7.080, R² 0.193
Fold 3 | Linear Regression -> RMSE 5.648, R² 0.360
Fold 4 | Linear Regression -> RMSE 4.639, R² 0.449
Fold 5 | Linear Regression -> RMSE 4.142, R² 0.405

Linear Regression Cross-Validation RMSE: 5.315 ± 1.012
Linear Regression Cross-Validation R²: 0.341 ± 0.089
```

5-Fold CV for k-Nearest Neighbors

Numeric features were scaled with StandardScaler(). Used k-Fold cross validation (k = 5) to find the best value for the choice of parameter k in k-Nearest Neighbors.

Results:

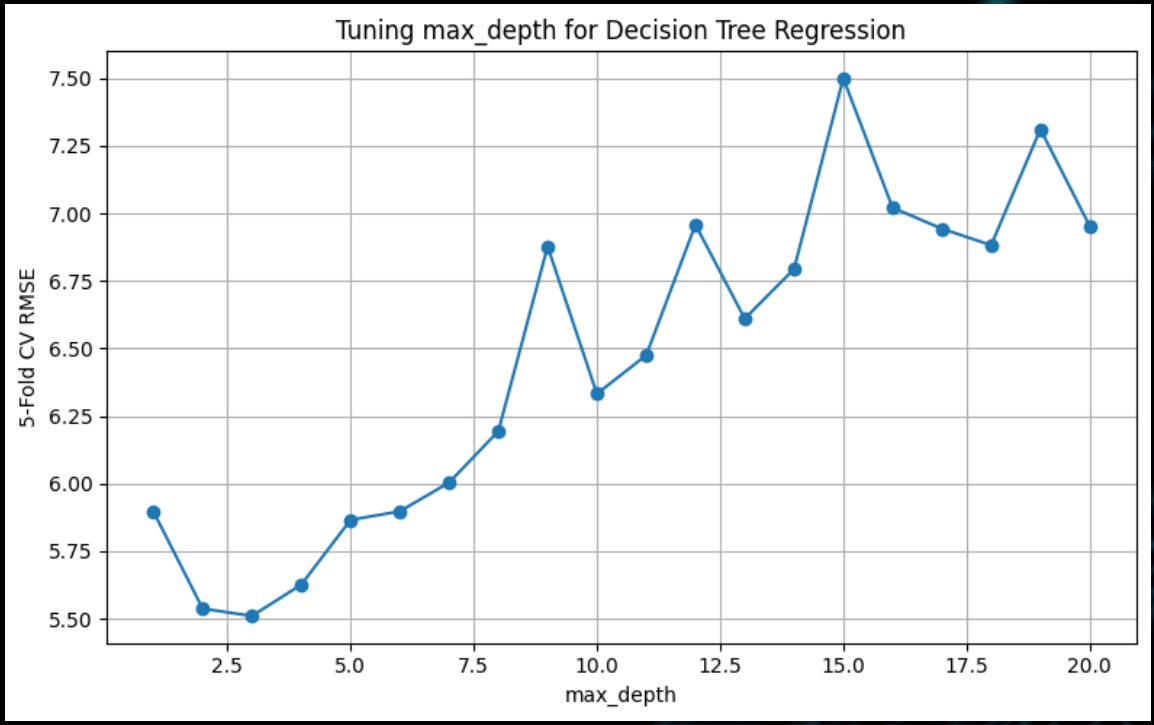



```
Best k: 20 with Cross-Validation RMSE = 5.306
```

5-Fold CV for Decision Tree

No scaling was done. Used k-Fold cross validation (k = 5) to find the best value for the choice of parameter max_depth.

Results:




```
Best max_depth: 3 with Cross-Validation RMSE = 5.511
```

5-Fold CV for Random Forest

No scaling was done. Used k-Fold cross validation (k = 5) to find the best values for the choices of parameters max_depth and n_estimators.

Results:

```
max_depth=3, n_estimators=10 -> Cross-Validation RMSE: 5.367
max_depth=3, n_estimators=50 -> Cross-Validation RMSE: 5.337
max_depth=3, n_estimators=100 -> Cross-Validation RMSE: 5.339
max_depth=5, n_estimators=10 -> Cross-Validation RMSE: 5.384
max_depth=5, n_estimators=50 -> Cross-Validation RMSE: 5.304
max_depth=5, n_estimators=100 -> Cross-Validation RMSE: 5.305
max_depth=10, n_estimators=10 -> Cross-Validation RMSE: 5.461
max_depth=10, n_estimators=50 -> Cross-Validation RMSE: 5.359
max_depth=10, n_estimators=100 -> Cross-Validation RMSE: 5.326
max_depth=15, n_estimators=10 -> Cross-Validation RMSE: 5.600
max_depth=15, n_estimators=50 -> Cross-Validation RMSE: 5.372
max_depth=15, n_estimators=100 -> Cross-Validation RMSE: 5.345
Best max_depth: 5, Best n_estimators: 50, Best Cross-Validation RMSE: 5.304
```

5-Fold CV for XGBoost

No scaling was done. Split the dataset into train vs. test (80/20) and ran a GridSearchCV on the training set only to find the best values for the choices of parameters n_estimators, max_depth, learning_rate and subsample.

Results:

Best Parameters (CV): {'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 50, 'subsample': 1.0}

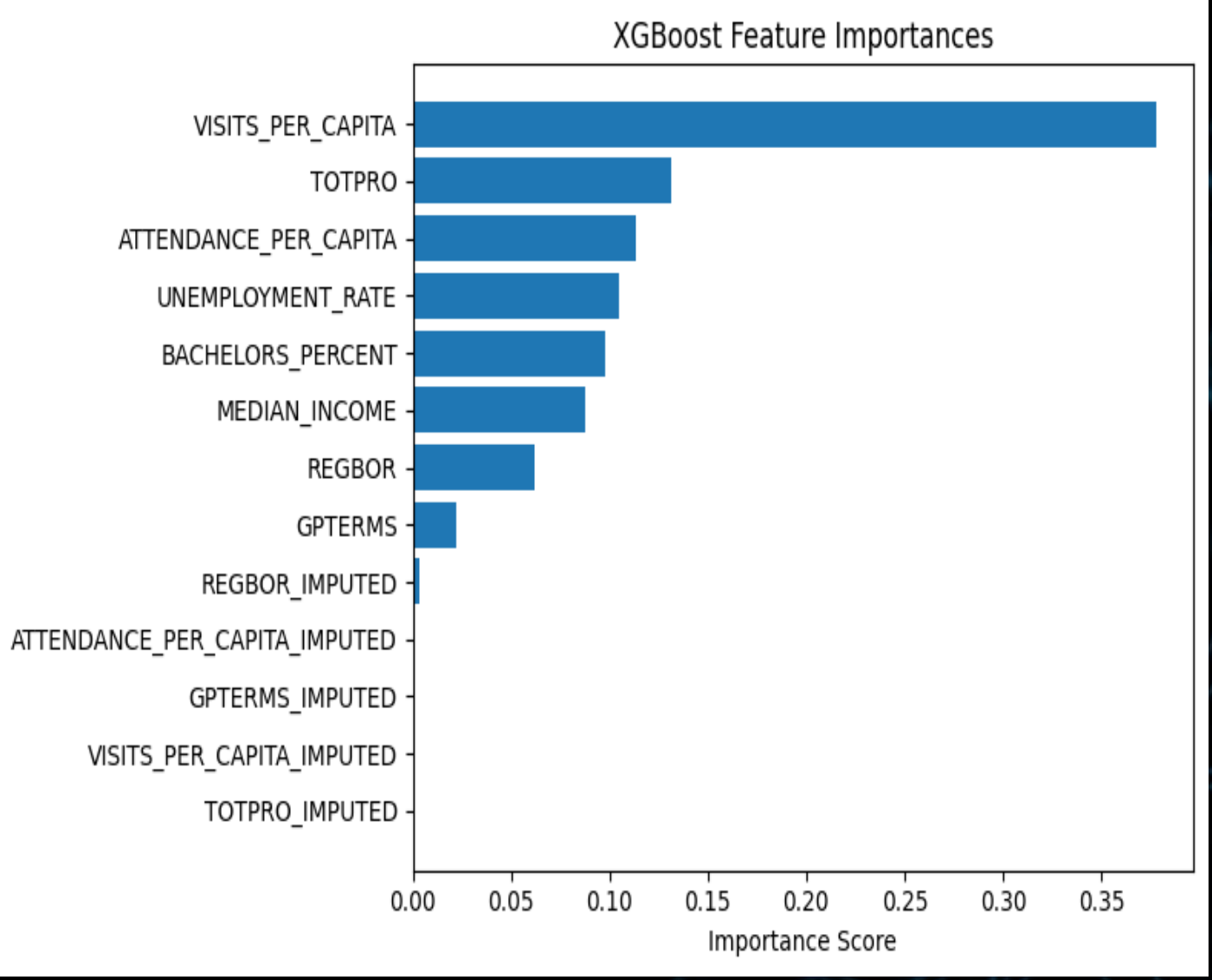
Best 5-Fold CV RMSE: 5.209

Since XGBoost has the lowest RMSE, it was retrained on full training data with the above hyperparameters and evaluated on held-out test set, whereby the results are:

`Test RMSE: 5.041, Test R²: 0.306`

CV RMSE 5.209 vs. Test RMSE 5.041 indicates the model generalizes slightly better to the unseen test set, meaning there is no severe overfitting. R^2 = 0.306 means that the selected features explain about 31% of the variance in checkouts per capita across ZIP codes, but there's still a fair amount of unexplained variation possibly due to some limitations about the dataset.

Feature Importances



1. Visits per Capita (38%)

·Top predictor by far, more frequent library visits translates directly into more checkouts.

·Reinforces the importance of outreach and programming that gets people through the doors.

2. Program Offerings (TOTPRO, 13%) & Attendance (11%)

·Both the *quantity* of synchronous programs and how well they're attended matter, libraries with richer, well-attended programming see higher checkouts.

·Suggests investments in engaging events pay dividends in overall circulation.

3. Socioeconomic Context (≈29% total)

·Unemployment Rate (10%), Bachelor's % (10%), Median Income (9%) together form nearly a third of the model's decision power.

·Confirms that community wealth, education level, and economic health still play a significant role in how much people check out.

4. Registered Users (6%) & Computers (2%)

·Having more registered cardholders and public internet terminals helps, but to a lesser extent.

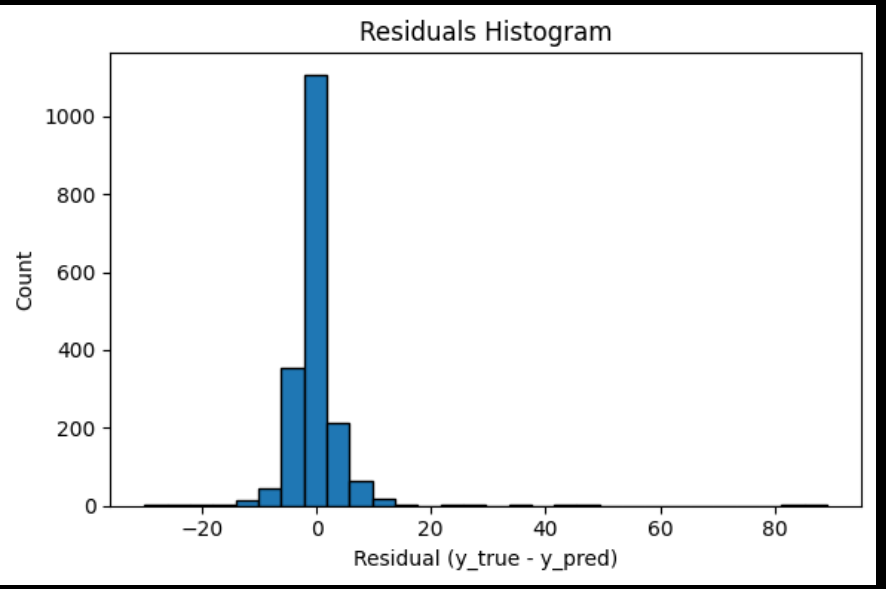
·These baseline infrastructure metrics are supportive, not drivers.

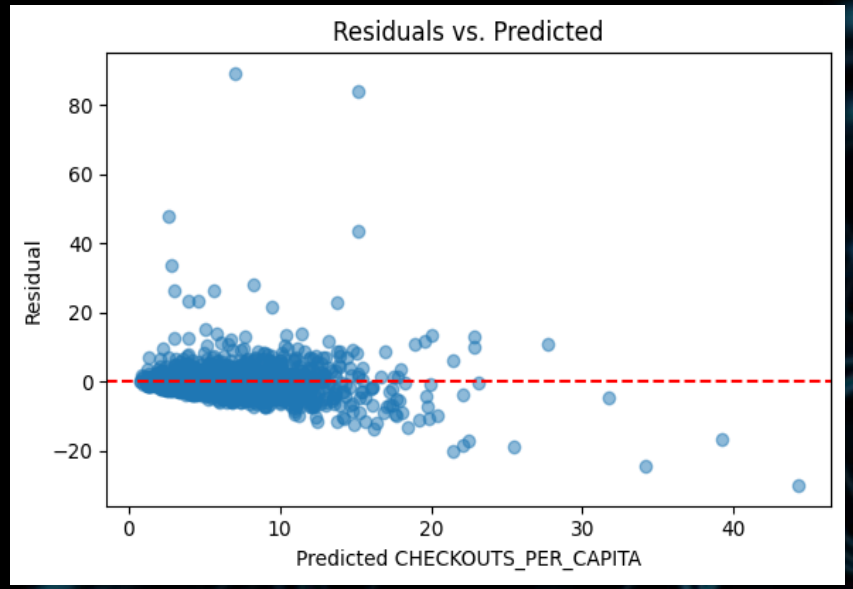
5. Imputation Flags (~0%)

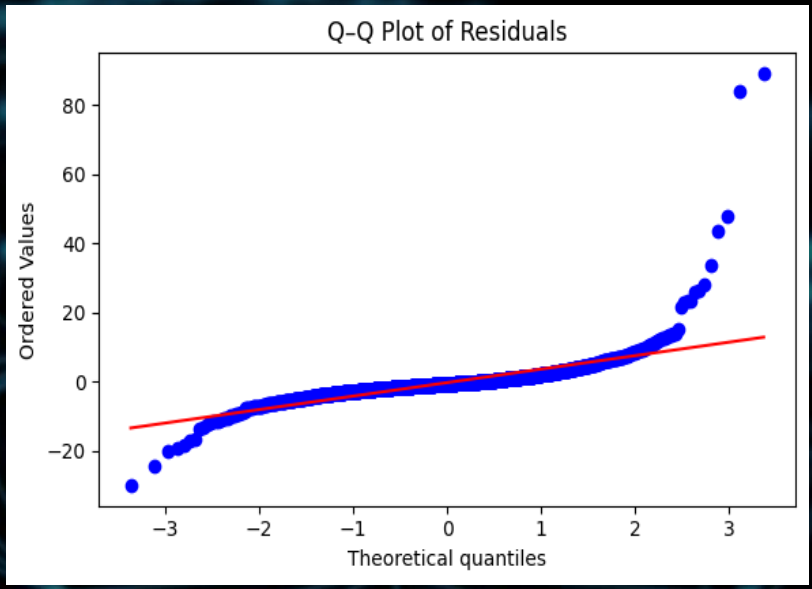
·The model ignored the "was this value imputed?" indicators.

·The cleaning and imputation procedures did not introduce spurious signals.

Residual Analysis







Residuals are largely symmetric and concentrated near zero, indicating unbiased predictions, though a few outliers suggest the model may underperform in extreme cases.

The residuals exhibit roughly constant spread across most predicted values, showing no major heteroskedasticity, although a handful of extreme ZIP codes at the low end are over-predicted and at the high end under-predicted.

Residuals are roughly normal in the center, although the deviations in the tails indicate a handful of extreme outliers consistent with our histogram and residual-vs-predicted plots.

Dataset Limitations

1. ZCTA vs. ZIP Code Mismatch

· “ZIP” areas (ZCTAs) in the Census dataset are only approximations of USPS ZIP Codes. Some libraries may fall just outside the true ZCTA boundaries, introducing geographic noise.

2. Cross-Sectional Snapshot

· All data are from Fiscal Year 2022; we don’t capture seasonal or year-to-year trends in visits, programs, or checkouts.

3. Unobserved Local Factors

· We lack variables on facility attributes, marketing efforts, building conditions, transit & walkability and so on, which likely influence library usage but aren’t in the data.

4. Aggregation Bias

· By averaging everything at the ZIP-code level, we lose the story of individual neighborhoods. For example, one ZIP could include both a busy urban area and a quieter suburb, our numbers blend them together and hide those local differences.

5. Model Residual Outliers

· A few ZIPs with extreme check-out rates remain under- or over-predicted, suggesting the features don’t fully explain those atypical cases.

Possible Future Work

1. Incorporating Time Series

· Pulling in historical PLS + ACS data to model trends, seasonality, or the impact of events (e.g. pandemic recovery).

2. Spatial Modeling

· Using geospatial methods to account for neighboring ZIP influences and reduce ZCTA boundary noise.

3. Richer Feature Sets

· Adding data on library staff credentials, program types, local school enrollment and such to better explain outliers.

4. Improving Imputation & Granularity

· For cells with heavy imputation, exploring multiple imputation (more robust) or smaller geographic units (census tracts) where data quality permits to capture neighborhood-level differences and reduce the blending effect of large ZIP areas.

5. Causal Analysis

· Moving beyond prediction to causal inference, e.g. measuring the effect of opening new computer labs or adding evening hours.