Predicting Economic Development Through Urban Growth Patterns: A Data-Driven Approach to City Planning

Zoe Hightower, Yve Hu, and Tylan Joshi

Abstract

This study aims to provide a predictive model, created using fundamental linear algebra and probabilistic topics, that explores how different urban development strategies–densification, increased public transportation, walkability, building expansion, or traffic congestion, affect economic growth, using GDP as a measurement of economic status. Using data from the 30 largest Standard Metropolitan Statistical Areas (SMSAs), a linear regression model is utilized to predict the GDP of each specific city by creating variables for the development strategies mentioned above. The results of the model aim to provide a sense of how each developmental strategy affects the economy. Model validation was found by seeing the correlation between the predicted and actual GDP of each city by calculating Spearman's Rank correlation coefficients. To prevent multicollinearity, a Pearson's correlation coefficient was run through each pair of variables. If variables were found to be statistically significantly similar ($p \geq 0.8$), a simultaneous variable was made using principal component analysis. Limitations can be found in the fact only the largest cities were analyzed, meaning rural areas may not have data that fit the model. This research aims to provide a model that will highlight the importance of different development patterns economically, helping cities better predict how different developments will affect the economy.

1 Introduction

1.1 Research Question

Throughout our exploration, we aim to answer the following question: "Can we predict how different urban development patterns will affect economic developments, and if so, can we create a valid predictive model to do so?"

1.2 Motivation and Background

The United States' urban development patterns since the 1930s have emphasized car dependence, build outwards from central cities, and restricting public funds of and limiting public transportation. Before the 1930s, the majority of Americans did not own cars. Most Americans lived in areas where daily tasks could be completed without the car. As a result, American urban areas were walkable, had extensive public transportation, and were designed in a manner such that there is a central area where business and government services occur. However, after the 1930s, the mass production of the automobile led to the American car ownership rate skyrocketing. This phenomenon meant that Americans no longer needed to live in central urban locations or use public transportation to benefit from the economic opportunities urban areas offer. "[T]he U.S. urban landscape resulted from a combination of car purchases, large public investments in road infrastructure, limited public investment in central cities, the existence of much population heterogeneity within cities and low cultural barriers to house (Nechyba and Walsh 185)." Thus, many Americans moved to suburbs like Levittown in Long Island, NY (i.e., one of the first model American suburbs that featured solely mass-produced single-family homes, parking lots, and no public transportation). Alongside this migration pattern, American public transportation networks were curtailed, such as the American cities' extensive streetcar and rail network. Likewise, the National Interstate and Defense Highways Act of 1956 further accelerated these new urban development patterns by establishing a federally-organized network of roads that allows for cars to be a convenient and preferable option of transportation. In addition, many cities implemented "urban renewal," where dense urban cores of cities were replaced by highways and parking lots. In the decades since, American urban areas have continued building outwards, where new residential, business, and other infrastructures are built far away from the historic city center—or "downtown." This process is known as "urban sprawl."

"Urban sprawl" has been correlated with multiple negative externalities, such as rising costs of living, commuting times, and environmental damage, and there is a strong correlation between economic cost and "urban sprawl" predictors. Similarly, urban areas that implement urban development that avoid "urban sprawl" (i.e., dense, non-car-dependent, public-transit-oriented, walkable, and where buildings are not spread out) have proven to benefit economically. Thus, the urban development patterns are predictors for a higher GDP per capita.

1.3 Variables

To investigate the association between patterns of urban growth and economic performance, this paper uses regional GDP (in billions, Y) as the explained variable and selects five indicators as explanatory variables for the model:

- Land-to-population Density (X_1): land density signifies how close together the population of an urban area is, and greater land density creates more human connectivity, which increases economic output.
- City Walkability (X_2): walkability allows for increased accessibility to places in the urban area for disabled individuals and individuals who cannot drive for various reasons, and increases the desirability of an urban area
- Public Transportation Ridership per Capita (X_3): increased public transportation usage reduces the risk of transportation-related accidents and associated economic costs, and reduces car-related cost for the population
- Building Compactness Index (X_4): building compactness allows for shorter distances and times to go from one part of an urban area to another part, which cuts down the economic cost of sprawl, regardless of the transportation option
- Traffic Congestion Index (X_5): reduced traffic congestion cuts commuting times, and more commuting time has more economic cost

1.4 Limitations and Assumptions

Limitations include the score of the prediction model's data size, scope, uncalculated collinearity, and potential outliers. The model only analyzes the 30 biggest MSAs in the United States. Because we look into the largest MSAs, smaller MSAs may have correlations between infrastructure and economy that bigger MSAs. Furthermore, urban development patterns may not be as correlated in rural areas. Though we attempt to use 5 distinct variables, there may be other important variables that are currently excluded from this model. Uncalculated collinearity is also a potential issue. Though we verify that the variables chosen are significantly correlated through statistical calculations, since we only look into 30 cities, variables could still be collinear given a larger data set. Another problem is that the New York MSA may be an outlier because New York has a large population. Though we attempt to solve this problem by ranking the predictions when verifying the model, the MSA could potentially be altering the model.

2 Proposed Analysis

2.1 Linear Regression Model

Multiple Linear Regression is a method that constructs a linear regression that best fits a set of non-linear data (Yale, n.d.). An empirical analysis approach is employed to identify factors influencing regional GDP. Through regression analysis, we can study the relationship between independent and dependent variables and quantify these relationships by establishing a mathematical model. The general equation of a multiple linear regression model is given by:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon_i \tag{1}$$

(1) Represents the linear variation in the dependent variable Y due to changes in the n independent variables X. Specifically, Y is the vector of dependent variables, typically represented as

$$Y = \left[Y_1, Y_2, \ldots, Y_n\right]^T$$

where each Y_i represents the value of corresponding observation. The independent variable X_i forms an $m * (n+1)$ matrix, where m is the number of observations and n is the number of independent variables; Each row of this matrix represents the observation values of the i-th sample:

$$X = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,n} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{m,1} & X_{m,2} & \cdots & X_{m,n} \end{bmatrix} \tag{2}$$

β_0 is the constant or intercept term, $\beta_1, \beta_2, ..., \beta_n$ are the coefficients of the independent variables:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} \tag{3}$$

ϵ_i is an m-dimensional vector of random errors:

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{bmatrix} \tag{4}$$

This form of analysis assumes that there is no multicollinearity among the independent variables, meaning that the independent variables should not be highly correlated with each other. Therefore, multicollinearity diagnostics are necessary for the model. To ensure the model achieves the desired effect, the independent variables must have a significant impact on the dependent variable and exhibit a strong correlation with it.

2.2 Ordinary Least Squares

To estimate the coefficients of the multiple linear regression model, we utilized the Ordinary Least Squares (OLS) method. OLS minimizes the sum of squared residuals, ensuring that the predicted GDP values align as closely as possible with the observed data (Yale, n.d). The equation of OLS is as follows:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{5}$$

For how it pertains to our study, $\hat{\beta}$ is the vector of estimated regression coefficients $(\beta_0, \beta_1, \ldots, \beta_5)$, X is the matrix of independent variables, including a column of ones for the intercept term, and Y the vector of observed GDP values.

2.3 Pearson Correlation Coefficient

Prior to creating a linear regression model , we must ensure that all given variables are not correlated themselves. To reduce collinearity we calculated the Pearson correlation coefficient between every variable chosen. The Pearson correlation coefficient equation is as follows:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{6}$$

Between each pair of the five mentioned variables, we will denote one as x and the other as y. Thus, calculating the correlation coefficient r, where x_i are the values of x, $\bar{x}$ is the mean of all x values, y_i are the values of y, and $\bar{y}$ is the mean of all y values. If the calculated r comes out to be such that $r \geq 0.8$ or $r \leq -0.8$, the two variables must then be condensed down to one singular variable. Otherwise, our data has significant collinearity. This is because variables with r values less than -0.8 or greater than 0.8 are said to be significantly correlated.

2.4 Principle Component Analysis

If it is found that two variables have collinearity (significantly correlated), we will perform Principal Component Analysis (PCA). PCA is a pedicure that uses orthogonal transformation to transform variables classified as correlated to each other into a set of uncorrelated variables. The goal is to reduce collinearity. The steps of PCA are as follows:

2.4.1 Step 1: Standardize

Standardize the data so that the mean of each variable is 0, and the standard deviation is 1. This is done by calculating the following equation for each input:

$$Z = \frac{X - \mu}{\sigma} \tag{7}$$

For each input, Z is the calculated standardized value, X is the original value, μ is the mean of that input's variable, and σ is the standard deviation of that variable.

2.4.2 Step 2: Covariance Matrix

Find the covariance matrix for the two variables using the following equation, where x_1 and x_2 are the two variables, $x_1^{(i)}$ and $x_2^{(i)}$ are the specific points, $\bar{x}_1$ and $\bar{x}_2$ are the means of the variables, and n is the number of data points:

$$\operatorname{cov}(x_1, x_2) = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_1^{(i)} - \bar{x}_1 \right) \left(x_2^{(i)} - \bar{x}_2 \right) \tag{8}$$

2.4.3 Step 3: Eigenvalues and Eigenvectors

We will then find the Covariance Matrix's eigenvalues and eigenvectors using the following equation:

$$A\mathbf{v} = \lambda\mathbf{v}$$

To find the eigenvalues, we solve the characteristic equation by setting the determinant of $(\lambda\ I - A) = 0$

Once the eigenvalues λ are found, we can substitute each eigenvalue back into the equation $(\lambda I - A)v = 0$ to find the corresponding eigenvectors $\mathbf{v}$.

2.4.4 Step 4: Principal Components

Find principal components by ordering found eigenvalues in descending order. The first in this order is the first principal component and the last is the nth principal component. We will choose a k amount of Eigenvalues and then form a matrix X of their eigenvectors.

2.4.5 Step 5: Transform

Multiply the standardized data matrix by the matrix X to a lower-dimensional representation of the data.

2.5 Spearman Rank Correlation Coefficient

After collecting the data, we want to validate whether the results can be labeled as accurately representing the data. This will be done by calculating the Spearman Rank Correlation Coefficient. This is a statistical calculation that first ranks the values of two given sets and then calculates whether those values have a statistically strong correlation. This validates the model, because the ranking aspect will allow there to be variation, since larger metropolitan areas on the line such as New York may be seen as outliers due to their large population. To calculate the Spearman Rank correlation, the following formula is used where d is the difference between the two ranks and n is the number of data points:

$$r = 1 - \frac{6 \sum d^2}{n(n-1)} \tag{9}$$

In order for our model to be classified as a strong predictive representation of the effect of the different factors on the economy, we must get an r value greater than 0.7 ($r \geq 0.7$).

3 Implementation

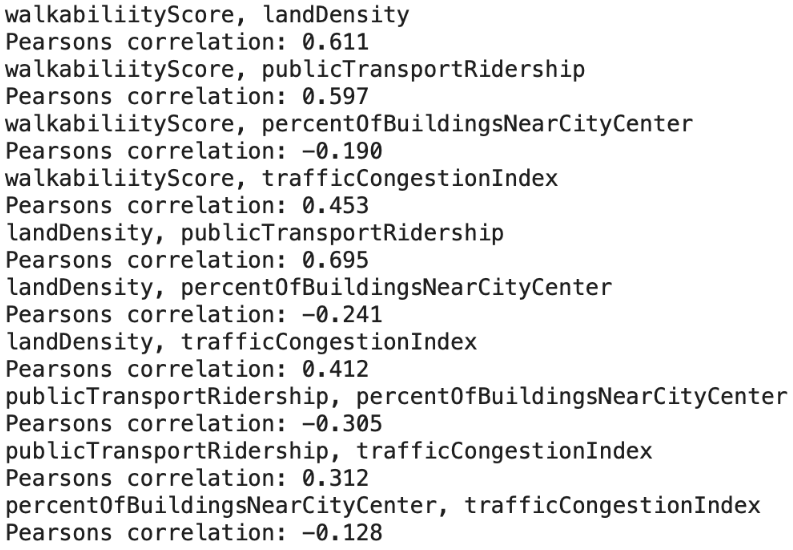
3.1 Step 1: Data Collection

This study uses 2020-2022 data from the 30 largest Standard Metropolitan Statistical Areas (SMSAs) in the United States as a sample. Data was taken between these years since the most recent census was 2020, meaning data on persons in that year is most accurate, and because some data in 2020 was not available, nearby years were chosen. Data was collected directly from first-party sources and journals: Census Reporter, Statista, American Public Transportation Association (APTA), and other reliable sources.

3.2 Step 2: Test Variable Correlations

One important assumption in a linear regression model is that each predictor does not have a high amount of correlation with other predictors. If multiple predictors are highly correlated, the standard error may be inflated, which may lead to inaccuracy in the model. As noted by C.H. Mason and W.D. Perreault, “Overall prediction is not affected, but interpretation of and conclusions based on the size of the regression coefficients, their standard errors, or the associated t-tests may be misleading because of the potentially confounding effects of collinearity. (Mason and Perreault, Jr. 268)” We used the Pearson Correlation Test to validate our model. Using a created Python code (see Listing 2) that implements the numPy and scipy frameworks, we observed the following results:

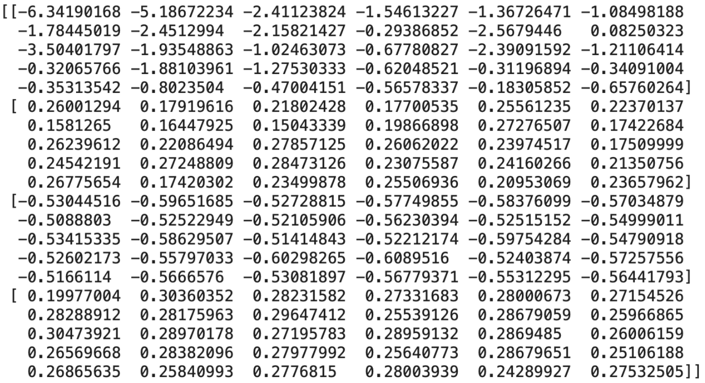
Table I: Results of Pearson Correlation Coefficients



```
walkabiliityScore, landDensity
Pearsons correlation: 0.611
walkabiliityScore, publicTransportRidership
Pearsons correlation: 0.597
walkabiliityScore, percentOfBuildingsNearCityCenter
Pearsons correlation: -0.190
walkabiliityScore, trafficCongestionIndex
Pearsons correlation: 0.453
landDensity, publicTransportRidership
Pearsons correlation: 0.695
landDensity, percentOfBuildingsNearCityCenter
Pearsons correlation: -0.241
landDensity, trafficCongestionIndex
Pearsons correlation: 0.412
publicTransportRidership, percentOfBuildingsNearCityCenter
Pearsons correlation: -0.305
publicTransportRidership, trafficCongestionIndex
Pearsons correlation: 0.312
percentOfBuildingsNearCityCenter, trafficCongestionIndex
Pearsons correlation: -0.128
```

Any two variables that have a number greater than or equal to 0.8 is considered to have a strong correlation, and any number greater than or equal to 0.5 is considered to have a moderate correlation in this field of study. However, the number depends on what academic field to which the correlation model relates: “The interpretation of r, however, is arbitrary depending on purpose and context. Hence, a value of r of 0.8 might be regarded as low in the physical sciences in which a physical law requires more rigorous verification but high in social and medical sciences in which a large number of individual X variables may be present. (Armstrong)” The strongest correlation between the predictor variables is “land density” and “public transport ridership”: at 0.695. This moderate correlation may be a result a need for more public transportation options if there is a higher population density. The second strongest correlation between the predictor variables is “walkability score” and “land density”: at 0.611. Similarly, this moderate correlation may be the result of the need for walkable infrastructure if there is a higher population density. Since there is a moderate collinearity, we used the Principal Component Analysis to reduce the dimensions of our data and remove the possible multicollinearity that is a result of possibly correlated predictor variables. We observed the following results:

Table II: PCA results



```
[[-6.34190168 -5.18672234 -2.41123824 -1.54613227 -1.36726471 -1.08498188
  -1.78445019 -2.4512994  -2.15821427 -0.29386852 -2.5679446   0.08250323
  -3.50401797 -1.93548863 -1.02463073 -0.67780827 -2.39091592 -1.21106414
  -0.32065766 -1.88103961 -1.27530333 -0.62048521 -0.31196894 -0.34091004
  -0.35313542 -0.8023504  -0.47004151 -0.56578337 -0.18305852 -0.65760264]
 [ 0.26001294  0.17919616  0.21802428  0.17700535  0.25561235  0.22370137
   0.1581265   0.16447925  0.15043339  0.19866898  0.27276507  0.17422684
   0.26239612  0.22086494  0.27857125  0.26062022  0.23974517  0.17509999
   0.24542191  0.27248809  0.28473126  0.23075587  0.24160266  0.21350756
   0.26775654  0.17420302  0.23499878  0.25506936  0.20953069  0.23657962]
 [-0.53044516 -0.59651685 -0.52728815 -0.57749855 -0.58376099 -0.57034879
  -0.5088803  -0.52522949 -0.52105906 -0.56230394 -0.52515152 -0.54999011
  -0.53415335 -0.58629507 -0.51414843 -0.52212174 -0.59754284 -0.54790918
  -0.52602173 -0.55797033 -0.60298265 -0.6089516  -0.52403874 -0.57257556
  -0.5166114  -0.5666576  -0.53081897 -0.56779371 -0.55312295 -0.56441793]
 [ 0.19977004  0.30360352  0.28231582  0.27331683  0.28000673  0.27154526
   0.28288912  0.28175963  0.29647412  0.25539126  0.28679059  0.25966865
   0.30473921  0.28970178  0.27195783  0.28959132  0.2869485   0.26006159
   0.26569668  0.28382096  0.27977992  0.25640773  0.28679651  0.25106188
   0.26865635  0.25840993  0.2776815   0.28003939  0.24289927  0.27532505]]
```

However, for our final prediction model, we chose to not use the Principle Component Analysis (PCA). We chose to use the data from PCA to have a better understanding of the relationship between variables, but decided to use the data we had set originally since no two variables had a strong correlation. In addition, PCA reduces dimensionality, some information is lost in the data. Furthermore, "PCA suffers from the fact that each principal component is a linear combination of all the original variables, thus it is often difficult to interpret the results. (Zou)"

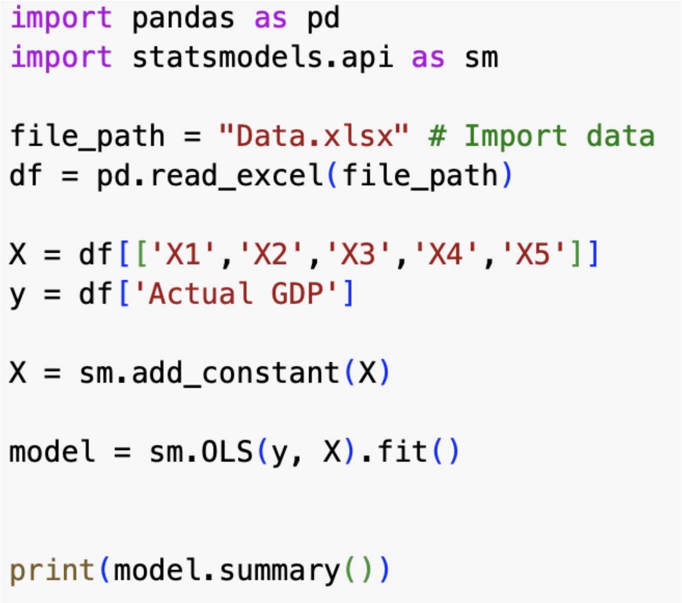
3.3 Step 3: Create Linear Regression Model

Based on the selected indicators, a multiple linear regression model is constructed with the GDP (Y) of metropolitan areas in the United States as the dependent variable. The independent variables are Land Density (X_1), Walkability Score (X_2), PT Ridership per Capita (X_3), Building Compactness Index (X_4), and Traffic Congestion Index (X_5):

$$\text{GDP} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon \tag{10}$$

First we calculated the Ordinary Least Squares (OLS). The OLS method ensures that the estimated coefficients provide the best linear unbiased estimator under the Gauss-Markov assumptions. The OLS method was implemented using Python's statsmodels library. The regression model was fitted using the selected independent variables (X_1, X_2, X_3, X_4, X_5) and the dependent variable - Actual GDP (Y). The Python implementation for calculating the OLS is as follows:

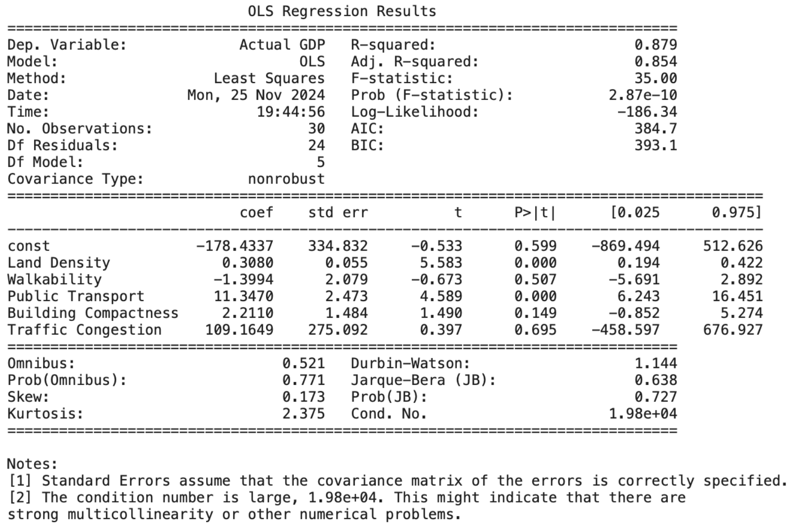
Table III: OLS code



```python
import pandas as pd
import statsmodels.api as sm

file_path = "Data.xlsx" # Import data
df = pd.read_excel(file_path)

X = df[['X1','X2','X3','X4','X5']]
y = df['Actual GDP']

X = sm.add_constant(X)

model = sm.OLS(y, X).fit()


print(model.summary())
```

Then we calculated the linear regression also using Python. The regression results are summarized below, providing key statistics for model evaluation and coefficient interpretation:

Table IV: Linear Regression and OLS Results



```
                            OLS Regression Results
==============================================================================
Dep. Variable:             Actual GDP   R-squared:                       0.879
Model:                            OLS   Adj. R-squared:                  0.854
Method:                 Least Squares   F-statistic:                     35.00
Date:                Mon, 25 Nov 2024   Prob (F-statistic):           2.87e-10
Time:                        19:44:56   Log-Likelihood:                -186.34
No. Observations:                  30   AIC:                             384.7
Df Residuals:                      24   BIC:                             393.1
Df Model:                           5
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
const                 -178.4337    334.832     -0.533      0.599    -869.494     512.626
Land Density             0.3080      0.055      5.583      0.000       0.194       0.422
Walkability             -1.3994      2.079     -0.673      0.507      -5.691       2.892
Public Transport        11.3470      2.473      4.589      0.000       6.243      16.451
Building Compactness     2.2110      1.484      1.490      0.149      -0.852       5.274
Traffic Congestion     109.1649    275.092      0.397      0.695    -458.597     676.927
==============================================================================
Omnibus:                        0.521   Durbin-Watson:                   1.144
Prob(Omnibus):                  0.771   Jarque-Bera (JB):                0.638
Skew:                           0.173   Prob(JB):                        0.727
Kurtosis:                       2.375   Cond. No.                     1.98e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.98e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Therefore our given model would be (rounded to 2 decimal cases):

$$\text{GDP} = -178.43 + 0.31X_1 - 1.40X_2 + 11.35X_3 + 2.21X_4 + 109.17X_5 \tag{11}$$

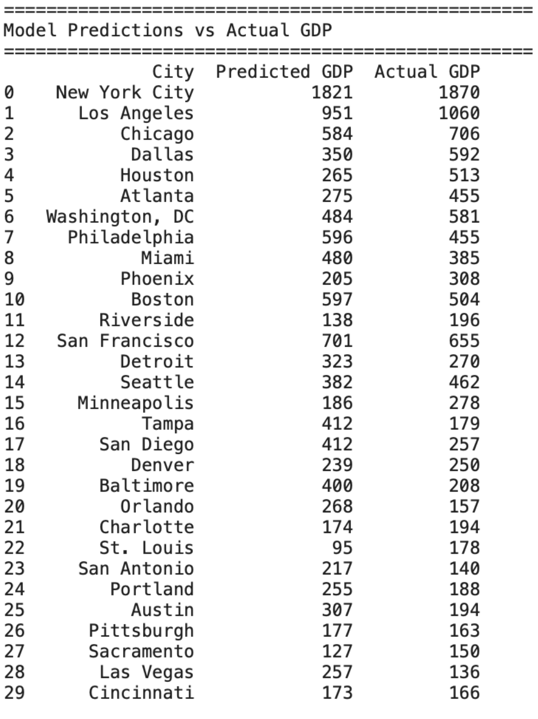
In particular, the R^2 value of 0.879 indicates that approximately 87.9% of the variance in GDP is accounted for, while the adjusted R^2 of 0.854 confirms a strong model fit after adjusting for predictors. The F-statistic (35.00, $p < 0.001$) shows the model is statistically significant.

The coefficient estimates show the relationships between urban development factors and GDP. Among the variables, Land Density (X_1) and Public Transport Ridership per Capita (X_3) have the most significant impact on GDP, with estimated coefficients and p-values of: 0.3080 ($p < 0.001$) and 11.3470 ($p < 0.001$), respectively.

4 Analysis

Using the calculated predictive model, we found the following predicted GDP for each MSA:

Table V: Predicted and Actual GDP

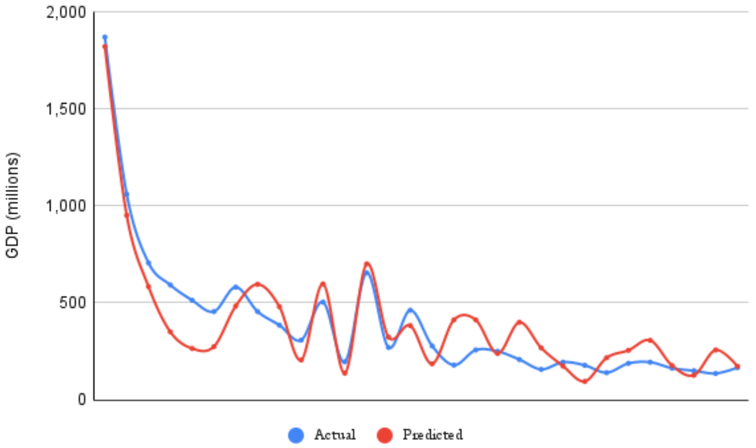


```
==================================================
Model Predictions vs Actual GDP
==================================================
              City  Predicted GDP  Actual GDP
0    New York City           1821        1870
1      Los Angeles            951        1060
2          Chicago            584         706
3           Dallas            350         592
4          Houston            265         513
5          Atlanta            275         455
6   Washington, DC            484         581
7     Philadelphia            596         455
8            Miami            480         385
9          Phoenix            205         308
10          Boston            597         504
11       Riverside            138         196
12   San Francisco            701         655
13         Detroit            323         270
14         Seattle            382         462
15     Minneapolis            186         278
16           Tampa            412         179
17       San Diego            412         257
18          Denver            239         250
19       Baltimore            400         208
20         Orlando            268         157
21       Charlotte            174         194
22       St. Louis             95         178
23     San Antonio            217         140
24        Portland            255         188
25          Austin            307         194
26      Pittsburgh            177         163
27      Sacramento            127         150
28       Las Vegas            257         136
29      Cincinnati            173         166
```

4.1 Visual Comparison

The above results were then graphed below. Looking at the graphical representations of the Predicted and Actual GDP values, it can be seen in the below line graph, that the numbers follow a very similar trend and are often similar values. Something noticed visually is that the model is more accurate on the mean of the data set, but in more extraneous points (the start of data or the last of the data).

Table VI: Line Graph of Predicted and Actual GDP



4.2 Statistically Comparing Predicted and Actual GDP

Since there visually appears to be a correlation between the predicted and actual GDP of the chosen MSA, we decided to calculate the Spearman Rank Correlation Coefficient. This method was chosen specifically because the aspect of ranking allows outliers, if any, to be discounted, since the general trend of rank would show more validity.

4.2.1 Spearman Rank Correlation Coefficient Results

Then we followed the Spearman's Rank Correlation Coefficient equation (equation 9). We first calculated the difference between the two results (d), then we calculated d^2 and it's respective sum. Plugging in these values to the equations (equation 9), we found that the Spearman Rank Correlation Coefficient (r) was approximately 0.73.

4.2.2 Discussion of Coefficient Results

According to statistical professionals, an r value of 0.73 would be classified as a strong positive correlation. This means that the relationship between actual and predicted GDP is strong. This statistically states that the model is quite accurate, but the positive notation, and the fact many of the predicted values are higher than the actual ones, suggest that our model might be overestimating some predictions.

4.3 Comments on Validity of Model

After reviewing both the graphical representations of the model and it's statistical analysis, it can be said that our model is, on average, remotely accurate. The statistical analysis reveals that the model and the actual results are similar to a strong extent and the graph of the predicted and actual values agree with this notion of similar trend.

5 Conclusion

In conclusion, this project is important because it provides a potential prediction model for how cities can expand current urban development projects. Our current model uses linear regression to preserve non-linear datasets, such as urban development patterns–which are not linear–to obtain a best fit model for that non-linear data. Because linear regression is prone to errors, we were careful to validate our model with the Pearson Correlation Test and the Spearman Rank. The Pearson Correlation Test checks if there is high correlation between different variables. High correlation may lead to inaccuracy within the model. The Spearman Rank has a similar method to that of the Pearson's correlation coefficient, but ranks the variables inputs prior to the calculation; this allowed for the predicted and actual GDP to be compared in a way that was most accurate. Correlation coefficient, highlights how different predictor variables concerning urban development and planning can contribute to the overall economic growth of an urban metropolitan area, represented by GDP per capita. As a result, this model may be used to highlight ways to improve other areas related to urban development, such as quality of life and environmental preservation.

References

[1] Zou, H., Hastie, T., Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics, 15*(2), 265. Retrieved from http://www.jstor.org/stable/27594179

[2] Nechyba, T. J., Walsh, R. P. (2004). Urban sprawl. *The Journal of Economic Perspectives, 18*(4), 177–200. Retrieved from http://www.jstor.org/stable/3216798

[3] Mason, C. H., Perreault, W. D. (1991). Collinearity, power, and interpretation of multiple regression analysis. *Journal of Marketing Research, 28*(3), 268–280. https://doi.org/10.2307/3172863

[4] Armstrong, R. A. (2019). Should Pearson's correlation coefficient be avoided? *Ophthalmic Physiological Optics, 39*(5), 316–327. https://doi.org/10.1111/opo.12636

[5] American Public Transportation Association. (2020). Economic impact of public transportation investment. Retrieved November 16, 2024, from https://www.apta.com/wp-content/uploads/APTA-Economic-Impact-Public-Transit-2020.pdf

[6] Narula, S. C., Wellington, J. F. (1977). Prediction, linear regression and the minimum sum of relative errors. *Technometrics, 19*(2), 185–190. https://doi.org/10.2307/1268628

[7] Statista. (n.d.). Real gross domestic product (GDP) of the United States by metropolitan area. Retrieved November 27, 2024, from https://www.statista.com/statistics/248083/real-gross-domestic-product-gdp-of-the-united-states-by-metropolitan-area/

[8] Census Reporter. (n.d.). New York-Newark-Jersey City, NY-NJ metro area. Retrieved November 27, 2024, from https://censusreporter.org/profiles/31000US35620-new-york-newark-jersey-city-ny-nj-metro-area/

[9] Urban Stats. (n.d.). Compactness of metropolitan areas. Retrieved November 27, 2024, from https://urbanstats.org/statistic.html?statname=Compactness&article_type=MSA&start=1&amount=All&universe=USA

[10] Walk Score. (n.d.). Walkability and transit scores. Retrieved November 27, 2024, from https://www.walkscore.com

[11] Yale University. (n.d.). Linear regression notes. Retrieved November 27, 2024, from http://www.stat.yale.edu/Courses/1997-98/101/linreg.html

6 Appendix

6.1 Code

Listing 1: Linear Regression Model

```

\# Linear Regression Modelimport pandas as pdfrom sklearn.linear\_model import
    LinearRegression

\# 1. IMPORT DATA FROM "Data.xlsx"file\_path = "Data.xlsx" \# uploaded in Colabdf =
    pd.read\_excel(file\_path)
```

```
X = df[['Land Density', 'Walkability','Public Transport','Building Compactness','Traffic
    Congestion']]y = df['Actual GDP']

model = LinearRegression()model.fit(X, y)

intercept = model.intercept\_coefficients = model.coef\_

\# PREDICT GDP USING LINREG MODELdf['Predicted GDP'] =
    model.predict(X).round(0).astype(int)

# PRINT INTERCEPT AND COEFFICIENTSprint(f"Intercept ():
    \{intercept:.2f\}")print("Coefficients:")for feature, coef in zip(X.columns,
    coefficients):print(f"- \{feature\}: \{coef:.4f\}")

\# 5. PREDICT GDP USING LINREG MODELdf['Predicted GDP'] =
    model.predict(X).round(0).astype(int)

\# 6. PRINT PREDICTED GDP VS ACTUAL GDPprint("=" * 50)print("Model Predictions vs Actual
    GDP")print("=" * 50)print(df[['City', 'Predicted GDP', 'Actual GDP']])
```

Listing 2: Pearson Correlation Coefficient Principle Component Analysis

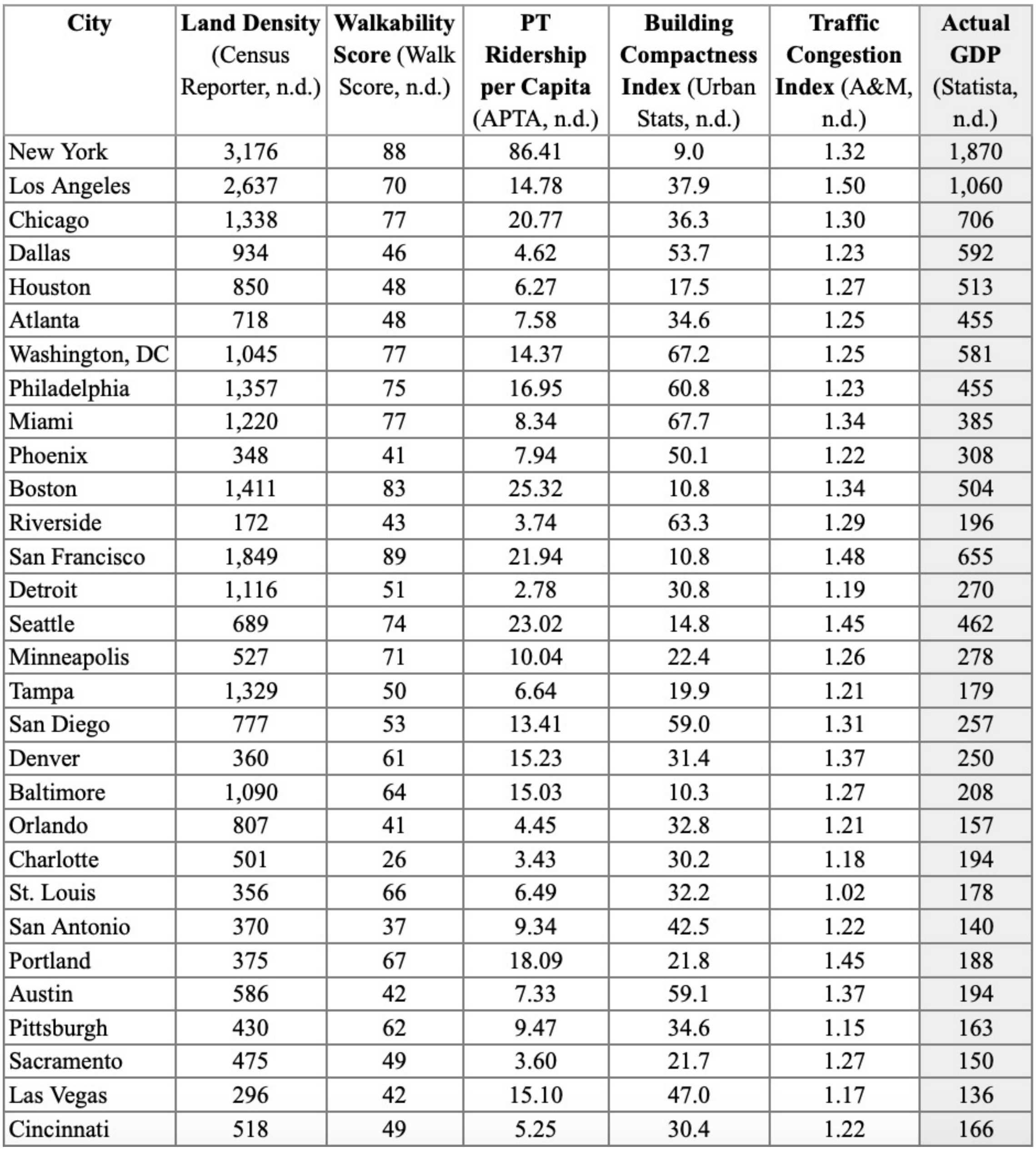
```
#Collinearity Test

from scipy.stats import pearsonr
from numpy import array
from scipy.linalg import svd
import numpy as np

walkabiliityScore = [88, 70, 77, 46, 48, 48, 77, 75, 77, 41, 83, 43, 89, 51, 74, 71, 50,
                     53, 61, 64, 41, 26, 66, 37, 67, 42, 62, 49, 42, 49]
landDensity = [3176, 2637, 1338, 934, 850, 718, 1045, 1357, 1220, 348, 1411, 172,
              1849, 1116, 689, 527, 1329, 777, 360, 1090, 807, 501, 356, 370,
              375, 586, 430, 475, 296, 518]
publicTransportRidership = [86.41, 14.78, 20.77, 4.62, 6.27, 7.58,
                          14.37, 16.96, 8.34, 7.94, 25.32, 3.74,
                          21.94, 2.78, 23.02, 10.04, 6.64, 13.41,
                          15.23, 15.03, 4.45, 3.43, 6.49, 9.34, 18.09,
                          7.33, 9.47, 3.6, 15.10, 5.25]
percentOfBuildingsNearCityCenter= [9.0, 37.9, 36.3, 53.7, 17.5, 34.6,
                                67.2, 60.8, 67.7, 50.1, 10.8, 63.3, 10.8,
                                30.8, 14.8, 22.4, 19.9, 59.0, 31.4, 10.3,
                                2.8, 30.2, 32.2, 42.5, 21.8, 59.1, 34.6,
                                21.7, 47.0, 30.4]
trafficCongestionIndex = [1.32, 1.50, 1.30, 1.23, 1.27, 1.25, 1.25, 1.23, 1.34, 1.22,
                        1.34, 1.29, 1.48, 1.19, 1.45, 1.26, 1.21, 1.31, 1.37, 1.27,
                        1.21, 1.18, 1.02, 1.22, 1.45, 1.37, 1.15, 1.27, 1.17, 1.22]

test1, _ = pearsonr(walkabiliityScore, landDensity)
print("walkabiliityScore, landDensity")
print('Pearsons correlation: %.3f' % test1)
test2, _ = pearsonr(walkabiliityScore, publicTransportRidership)
print("walkabiliityScore, publicTransportRidership")
print('Pearsons correlation: %.3f' % test2)
test3, _ = pearsonr(walkabiliityScore, percentOfBuildingsNearCityCenter)
print("walkabiliityScore, percentOfBuildingsNearCityCenter")
print('Pearsons correlation: %.3f' % test3)
test4, _ = pearsonr(walkabiliityScore, trafficCongestionIndex)
print("walkabiliityScore, trafficCongestionIndex")
print('Pearsons correlation: %.3f' % test4)
test5, _ = pearsonr(landDensity, publicTransportRidership)
print("landDensity, publicTransportRidership")
print('Pearsons correlation: %.3f' % test5)
test6, _ = pearsonr(landDensity, percentOfBuildingsNearCityCenter)
print("landDensity, percentOfBuildingsNearCityCenter")
print('Pearsons correlation: %.3f' % test6)
test7, _ = pearsonr(landDensity, trafficCongestionIndex )
print("landDensity, trafficCongestionIndex")
print('Pearsons correlation: %.3f' % test7)
test8, _ = pearsonr(publicTransportRidership, percentOfBuildingsNearCityCenter)
print("publicTransportRidership, percentOfBuildingsNearCityCenter")
print('Pearsons correlation: %.3f' % test8)
test9, _ = pearsonr(publicTransportRidership, trafficCongestionIndex)
print("publicTransportRidership, trafficCongestionIndex")
print('Pearsons correlation: %.3f' % test9)
test10, _ = pearsonr(percentOfBuildingsNearCityCenter, trafficCongestionIndex)
print("percentOfBuildingsNearCityCenter, trafficCongestionIndex")
print('Pearsons correlation: %.3f' % test10)

cities = np.array([walkabiliityScore, landDensity, publicTransportRidership,
    percentOfBuildingsNearCityCenter, trafficCongestionIndex]).T
cities_standardized = (cities - np.mean(cities, axis=None)) / np.std(cities, axis=None)
covariance_matrix = np.cov(cities_standardized.T)
eigenvalues, eigenvectors = np.linalg.eig(covariance_matrix)
eigenvectors = eigenvectors[:, np.argsort(eigenvalues)[::-1]]
cities_transformed = np.dot(cities_standardized, eigenvectors[:, :4])
print(cities_transformed.T)
```

Listing 3: Ordinary Least Squares

```
# OLS
import pandas as pd
import statsmodels.api as sm

file_path = "Data.xlsx" # Import data
df = pd.read_excel(file_path)

X = df[['Land Density', 'Walkability', 'Public Transport',
       'Building Compactness', 'Traffic Congestion']]
y = df['Actual GDP']

X = sm.add_constant(X)

model = sm.OLS(y, X).fit()


print(model.summary())
```

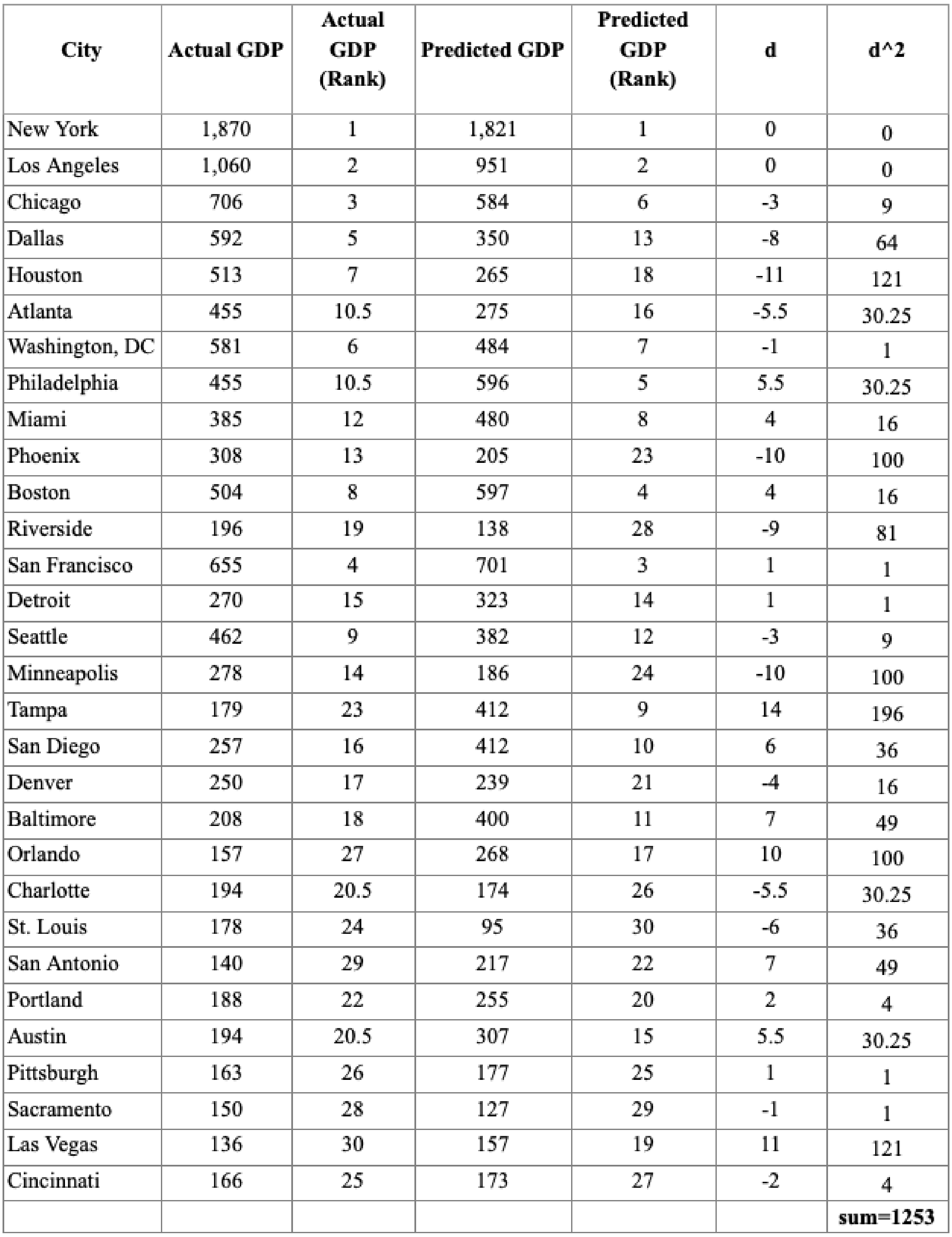
6.2 Data

Table VII: Raw Data



City	**Land Density** (Census Reporter, n.d.)	**Walkability Score** (Walk Score, n.d.)	**PT Ridership per Capita** (APTA, n.d.)	**Building Compactness Index** (Urban Stats, n.d.)	**Traffic Congestion Index** (A&M, n.d.)	**Actual GDP** (Statista, n.d.)
New York	3,176	88	86.41	9.0	1.32	1,870
Los Angeles	2,637	70	14.78	37.9	1.50	1,060
Chicago	1,338	77	20.77	36.3	1.30	706
Dallas	934	46	4.62	53.7	1.23	592
Houston	850	48	6.27	17.5	1.27	513
Atlanta	718	48	7.58	34.6	1.25	455
Washington, DC	1,045	77	14.37	67.2	1.25	581
Philadelphia	1,357	75	16.95	60.8	1.23	455
Miami	1,220	77	8.34	67.7	1.34	385
Phoenix	348	41	7.94	50.1	1.22	308
Boston	1,411	83	25.32	10.8	1.34	504
Riverside	172	43	3.74	63.3	1.29	196
San Francisco	1,849	89	21.94	10.8	1.48	655
Detroit	1,116	51	2.78	30.8	1.19	270
Seattle	689	74	23.02	14.8	1.45	462
Minneapolis	527	71	10.04	22.4	1.26	278
Tampa	1,329	50	6.64	19.9	1.21	179
San Diego	777	53	13.41	59.0	1.31	257
Denver	360	61	15.23	31.4	1.37	250
Baltimore	1,090	64	15.03	10.3	1.27	208
Orlando	807	41	4.45	32.8	1.21	157
Charlotte	501	26	3.43	30.2	1.18	194
St. Louis	356	66	6.49	32.2	1.02	178
San Antonio	370	37	9.34	42.5	1.22	140
Portland	375	67	18.09	21.8	1.45	188
Austin	586	42	7.33	59.1	1.37	194
Pittsburgh	430	62	9.47	34.6	1.15	163
Sacramento	475	49	3.60	21.7	1.27	150
Las Vegas	296	42	15.10	47.0	1.17	136
Cincinnati	518	49	5.25	30.4	1.22	166

Table VIII: Chart for Spearman's Rank Correlation Coefficient calculations



City	Actual GDP	Actual GDP (Rank)	Predicted GDP	Predicted GDP (Rank)	d	d^2
New York	1,870	1	1,821	1	0	0
Los Angeles	1,060	2	951	2	0	0
Chicago	706	3	584	6	-3	9
Dallas	592	5	350	13	-8	64
Houston	513	7	265	18	-11	121
Atlanta	455	10.5	275	16	-5.5	30.25
Washington, DC	581	6	484	7	-1	1
Philadelphia	455	10.5	596	5	5.5	30.25
Miami	385	12	480	8	4	16
Phoenix	308	13	205	23	-10	100
Boston	504	8	597	4	4	16
Riverside	196	19	138	28	-9	81
San Francisco	655	4	701	3	1	1
Detroit	270	15	323	14	1	1
Seattle	462	9	382	12	-3	9
Minneapolis	278	14	186	24	-10	100
Tampa	179	23	412	9	14	196
San Diego	257	16	412	10	6	36
Denver	250	17	239	21	-4	16
Baltimore	208	18	400	11	7	49
Orlando	157	27	268	17	10	100
Charlotte	194	20.5	174	26	-5.5	30.25
St. Louis	178	24	95	30	-6	36
San Antonio	140	29	217	22	7	49
Portland	188	22	255	20	2	4
Austin	194	20.5	307	15	5.5	30.25
Pittsburgh	163	26	177	25	1	1
Sacramento	150	28	127	29	-1	1
Las Vegas	136	30	157	19	11	121
Cincinnati	166	25	173	27	-2	4
						sum=1253