ANALYZING GEODESIC DISTANCE AS AN ESTIMATOR FOR ACCESSIBILITY

Implementing models to predict total travel duration between points in NYC

Abstract

Prior to the development of new public "goods" such as community hospitals, swimming pools or post offices, research is often conducted to maximize the accessibility of said goods to the public. This type of research has historically relied on the use of the geodesic distance computed between points to estimate accessibility. Today, companies can provide precise routes, and travel estimates, but these services are not free.

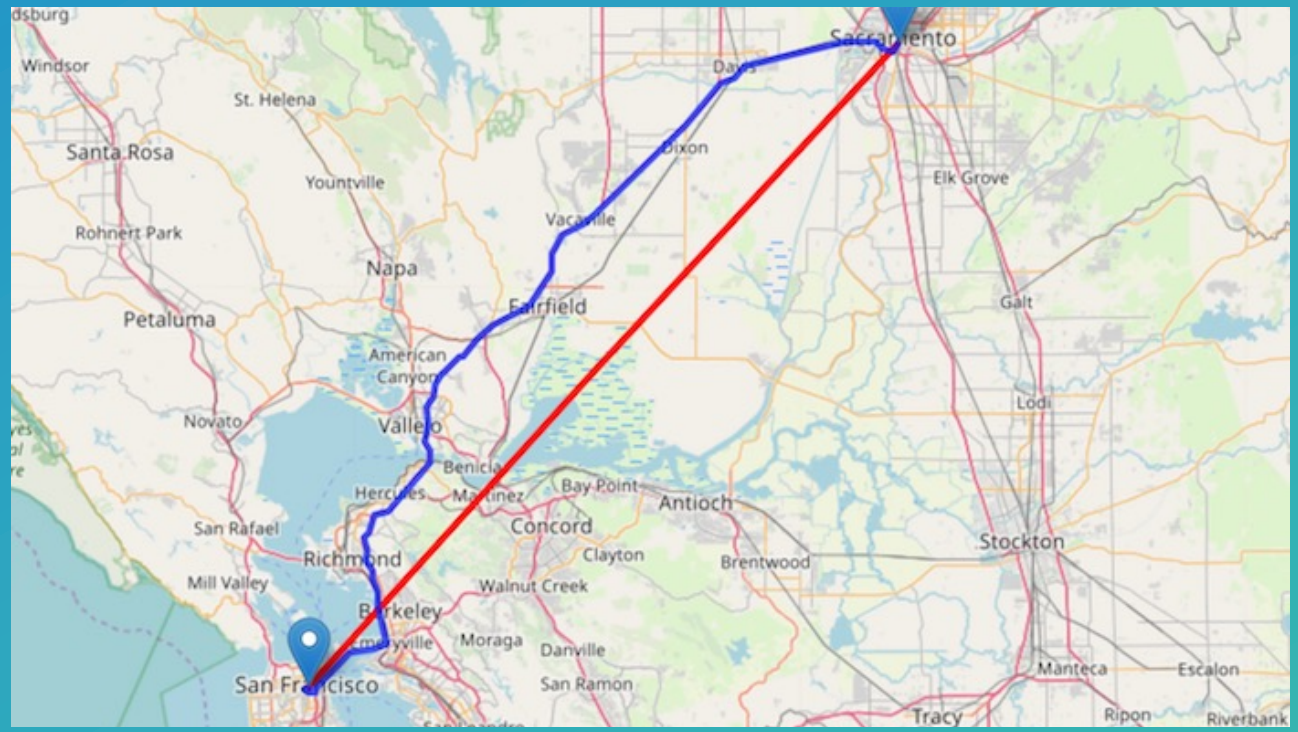
In this study, I analyze the relationship between the `google_duration` and the `geodesic_distance` between points in New York City to determine if `geodesic_distance` can still be reliably used as an estimator for travel time and accessibility. Datasets included in the analysis consist of:

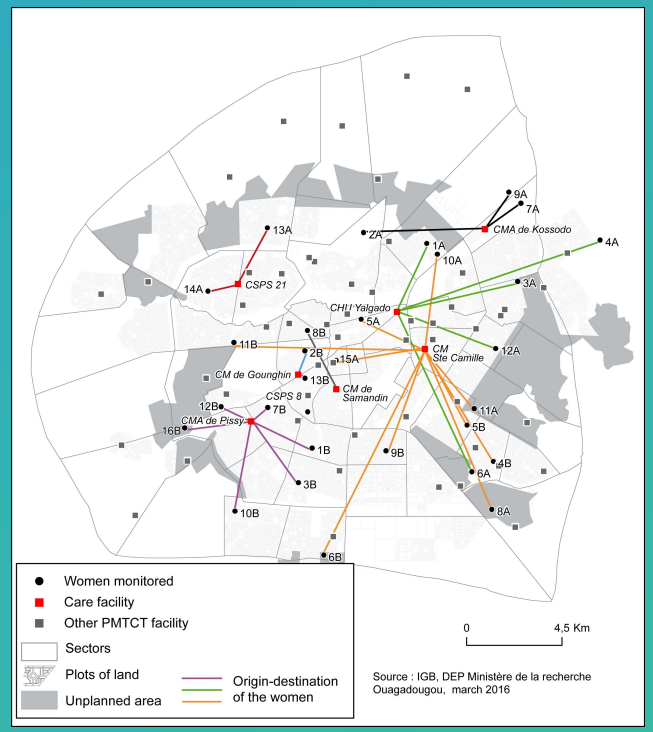
- Address Dataset:
  - 249 entries
  - All NYC-based US Post Offices and associated addresses
  - Sourced from postallocations.com
- Routes Dataset:
  - 30,876 entries
  - Route information for every combination of 2 addresses without repetition, which includes the `geodesic_distance`, `google_distance`, and `google_duration`
  - Manufactured from the Address Dataset using Python, Geopy, and Google Maps API

Results of the analysis show that there is a strong correlation between the `geodesic_distance` and `google_duration` between two points (0.87). Additionally, when implementing a polynomial-2 regression using `geodesic_distance` and other covariate variables, we achieve an Adjusted R^2 of 0.834 and a RMSE of 294 seconds (4.9 minutes). As expected, this model fails when there are significant obstacles between two points (rivers, JFK airport, etc.). The use of additional covariate variables could further improve this model's results.

The results of this study suggest that for non-critical public developments in NYC, the use of `geodesic_distance` for estimating accessibility can be considered.

STRAIGHT LINE DISTANCE
VS
COMPUTED ROUTE DISTANCE





Project Summary

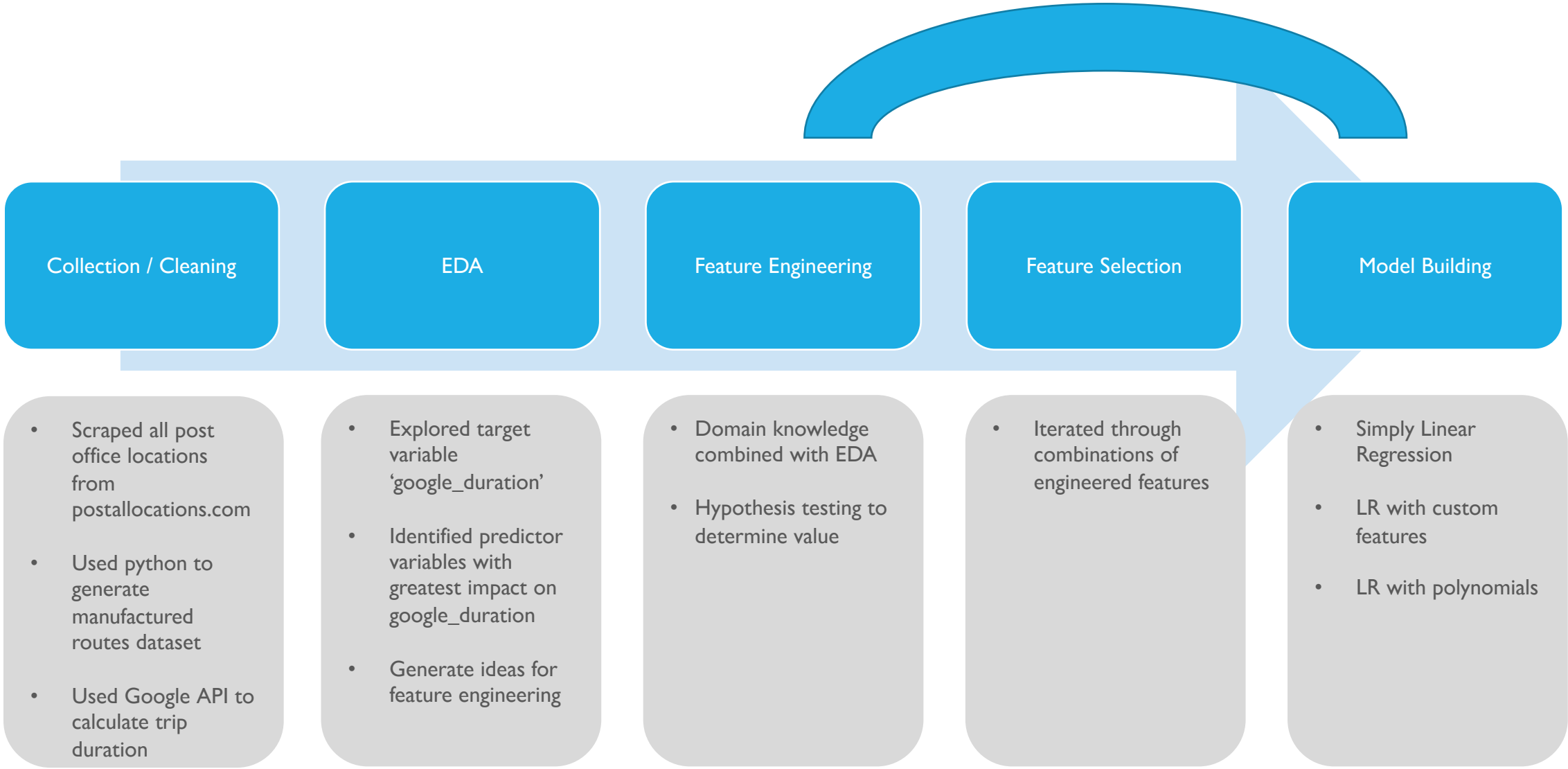
Data

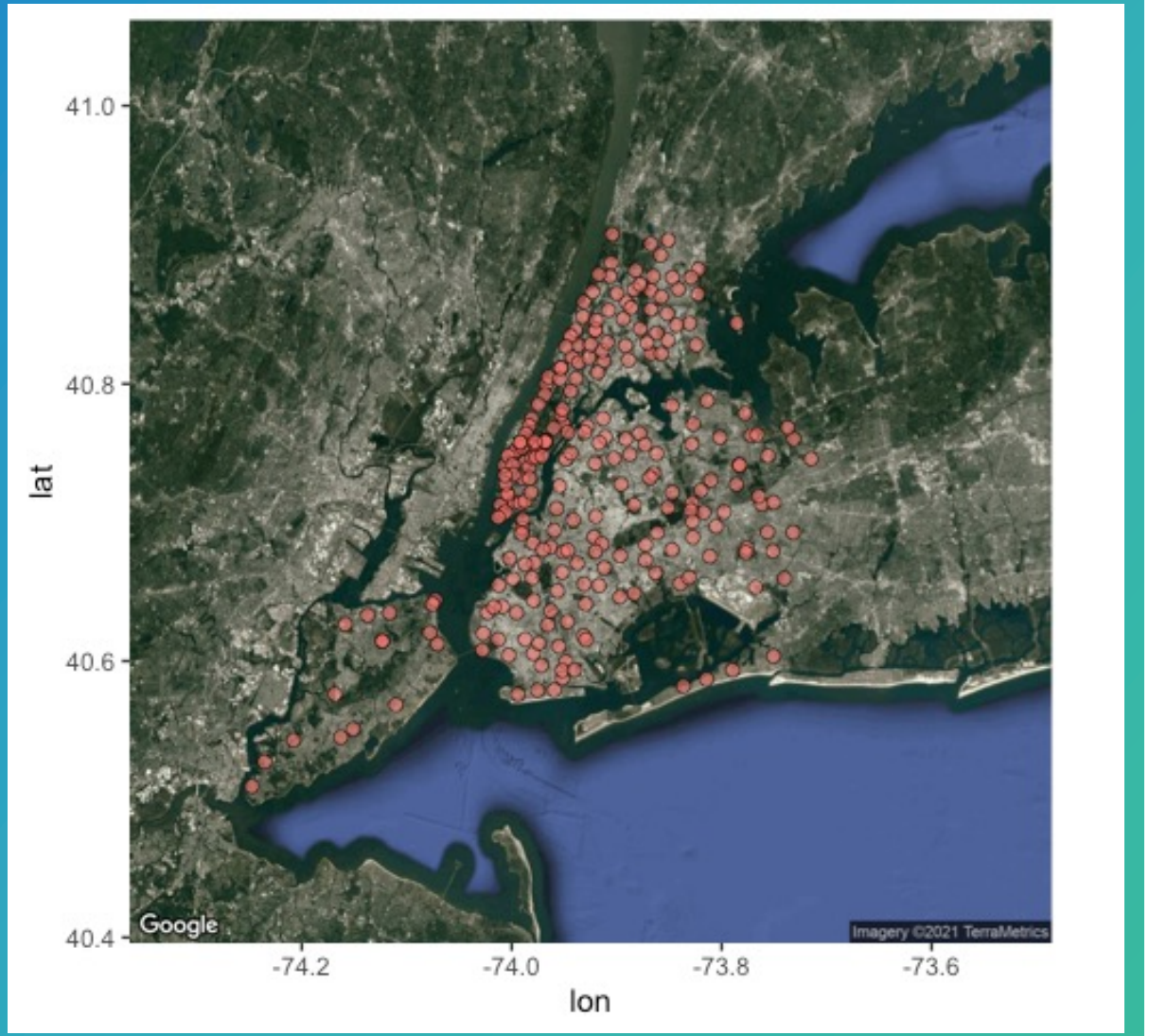
- postallocations.com: *all NYC post offices*
  - *249 locations in all 5 boroughs*
- manufactured data: *all routes between each post office*
  - *30k+ routes*
- Target Variable: *'google_duration' (time of trip)*

Objectives

- Build predictive regression model to estimate of total travel time given straight line distance
- Optimize Adjusted R^2 as primary performance metric

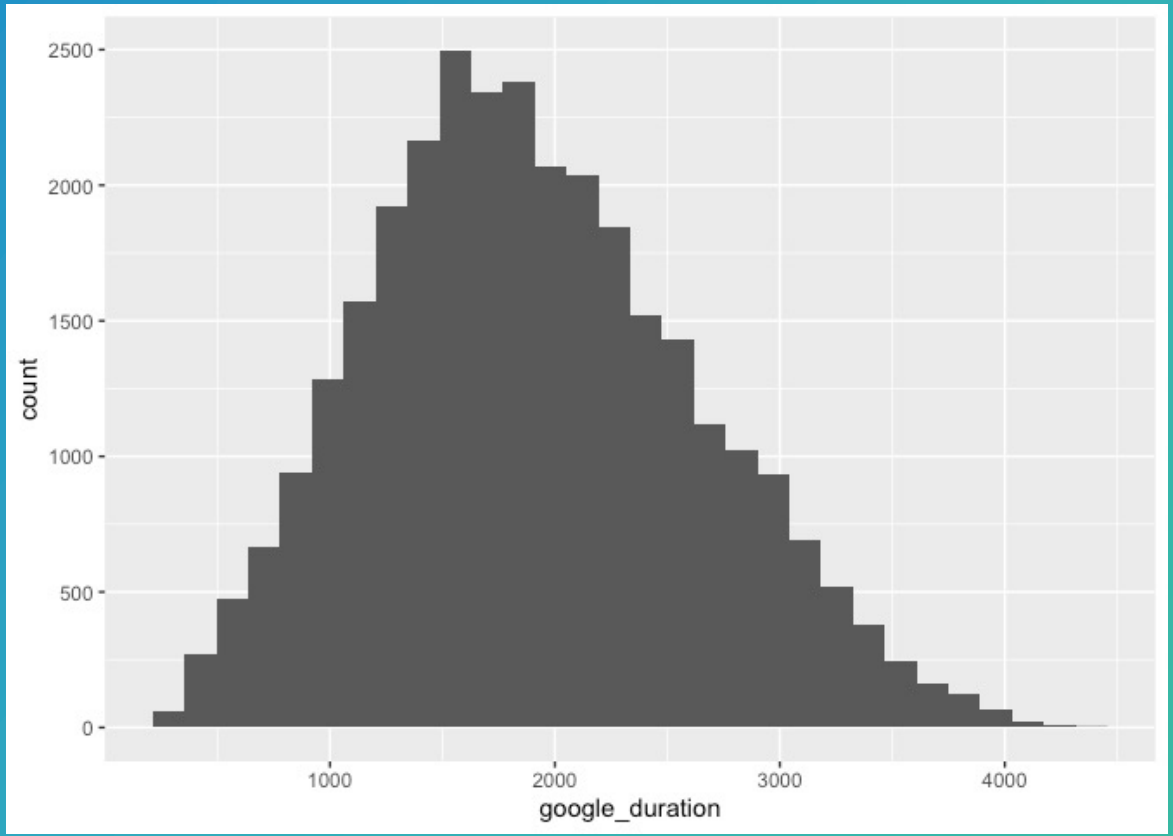
The Process


Collection / Cleaning
EDA
Feature Engineering
Feature Selection
Model Building
• Scraped all post office locations from postallocations.com
• Used python to generate manufactured routes dataset
• Used Google API to calculate trip duration
• Explored target variable 'google_duration'
• Identified predictor variables with greatest impact on google_duration
• Generate ideas for feature engineering
• Domain knowledge combined with EDA
• Hypothesis testing to determine value
• Iterated through combinations of engineered features
• Simply Linear Regression
• LR with custom features
• LR with polynomials



DATA: POST OFFICE LOCATIONS

- Initial dataset comprised of all (249) post office locations in NYC
- Post offices are placed based on population density and accessibility, and will typically be placed uniformly
- Equal number of post offices in all Boroughs except Staten Island

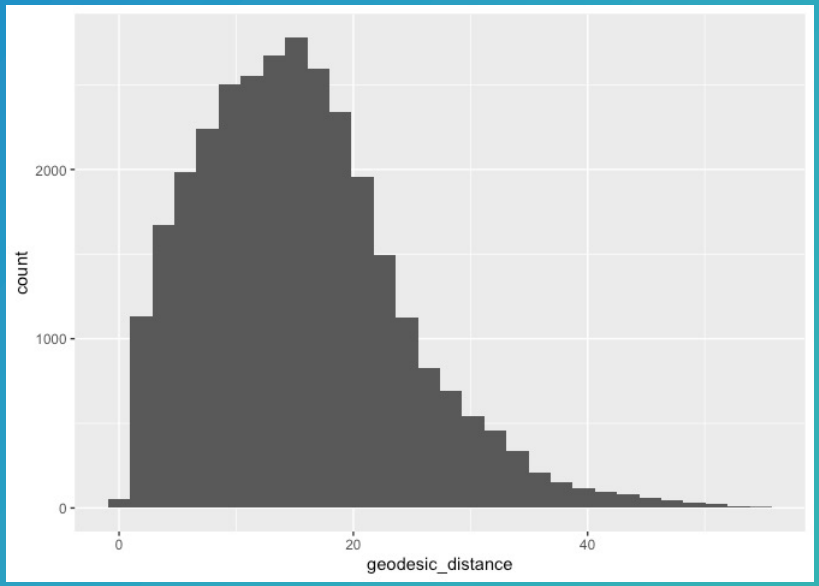


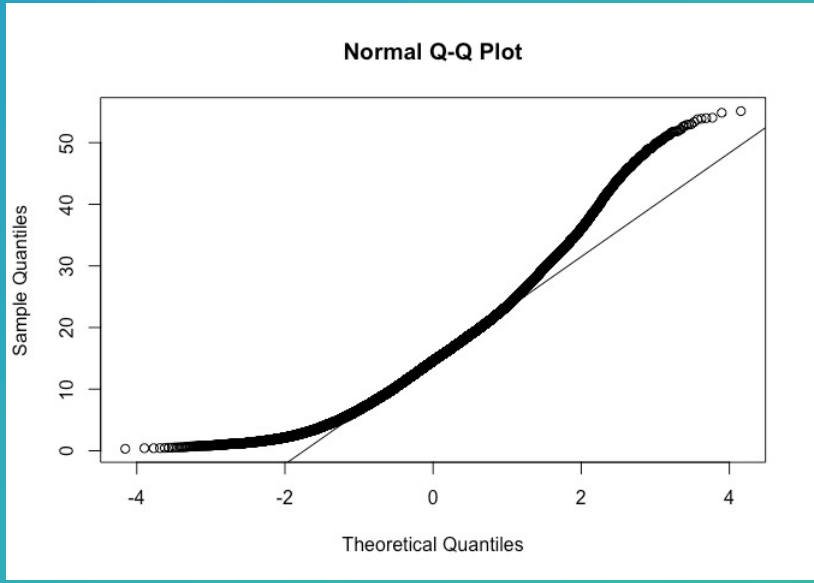
DATA: ROUTES DATASET

- The routes dataset is comprised of all (30k+) routes between pairs of post office locations

- For each route, we record the following:
  - google_duration
  - geodesic_distance
  - Starting borough
  - Ending borough
  - Start zipcode
  - End zipcode

- Response variable `google_duration` is normally distributed with a mean of 1901 seconds and a standard deviation of 723 seconds

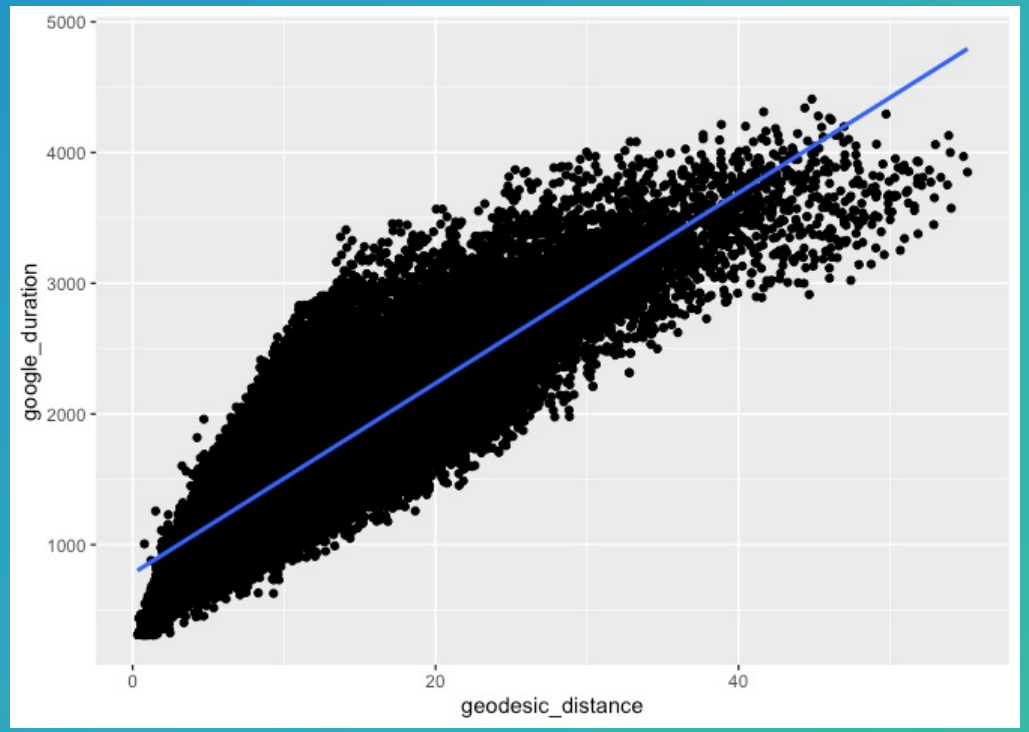




DATA: ROUTES DATASET

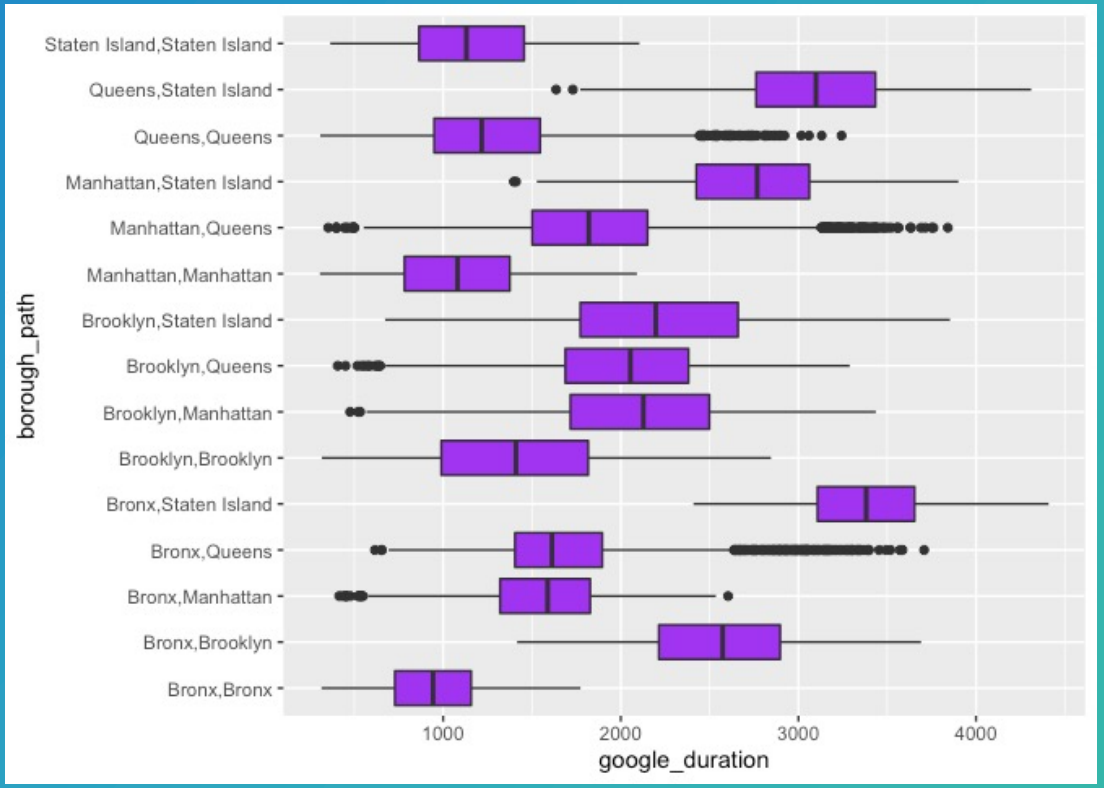
- Predictor variable “geodesic_distance” is not normally distributed.
- Shapiro Wilks test for normality produces p-value of 0.0016 when evaluating distribution
- Log transformation and Min-Max transformation fail to produce normality
- Considered removing outliers, but opted instead to leave them in

EDA – RESPONSE VS PREDICTOR



google_duration = 778.66 + 72.83 * **geodesic_distance**

- There is a clear relationship between primary predictor variable and response variable
- Correlation = 0.87
- Baseline linear model using no additional features with Adjusted R2 = .757
  - Intercept: 778.66
  - Slope: 72.83




```
anova <- aov(google_duration ~ borough_path, data = data)

summary(anova)
```

```
##                Df    Sum Sq   Mean Sq F value Pr(>F)
## borough_path    14 9.100e+09 649987339    2858 <2e-16 ***
## Residuals    30773 6.999e+09    227452
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

EDA – BOROUGHS

- A route's start and end borough provides additional information to our model
- Intuitively, routes that start and end in the same borough are typically shorter duration
- Any route that includes Staten Island tends to have a longer trip duration (bridges, etc).
- Hypothesis test confirms significance of boroughs (p-value = 2e^-16)

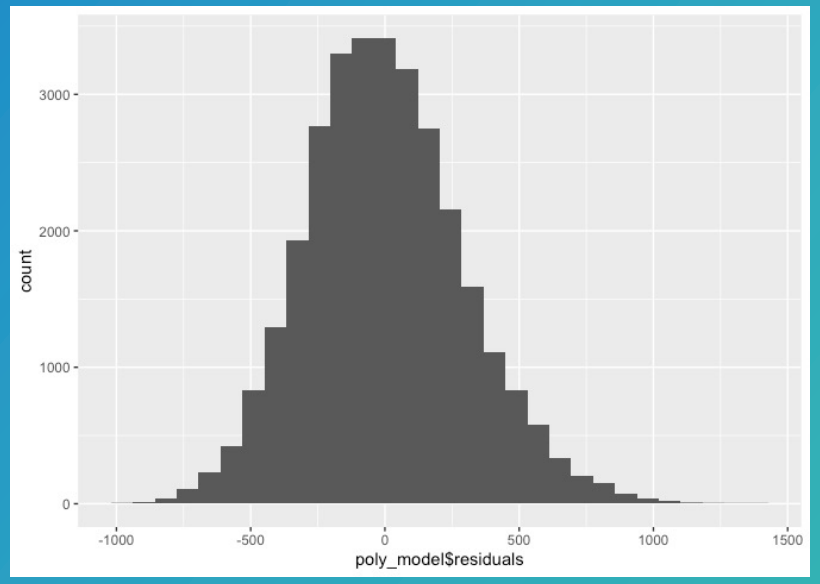
ADJUSTED R2:
0.83

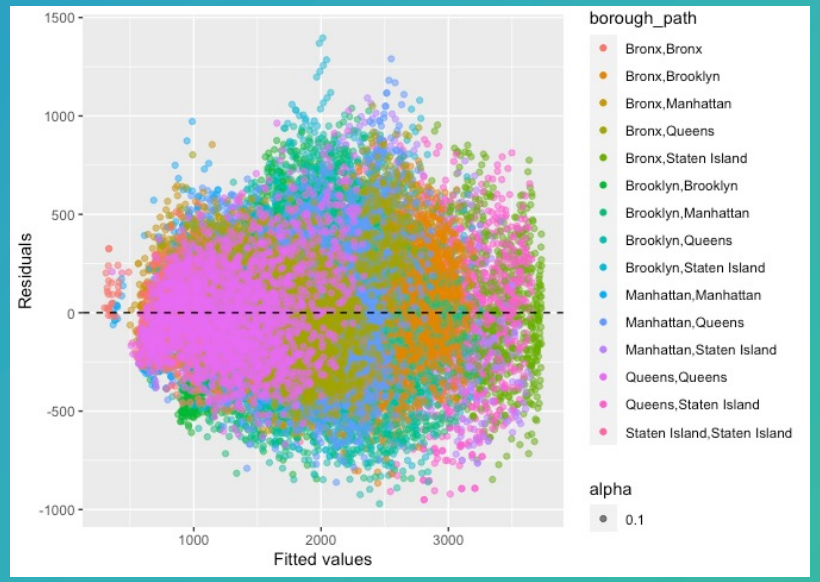
RSME
294 SECONDS

MODEL BUILDING

- Tested:
  - Log normalization vs non Log
  - 1 vs 2-polynomials
  - Covariates (boroughs and zips) vs no covariates

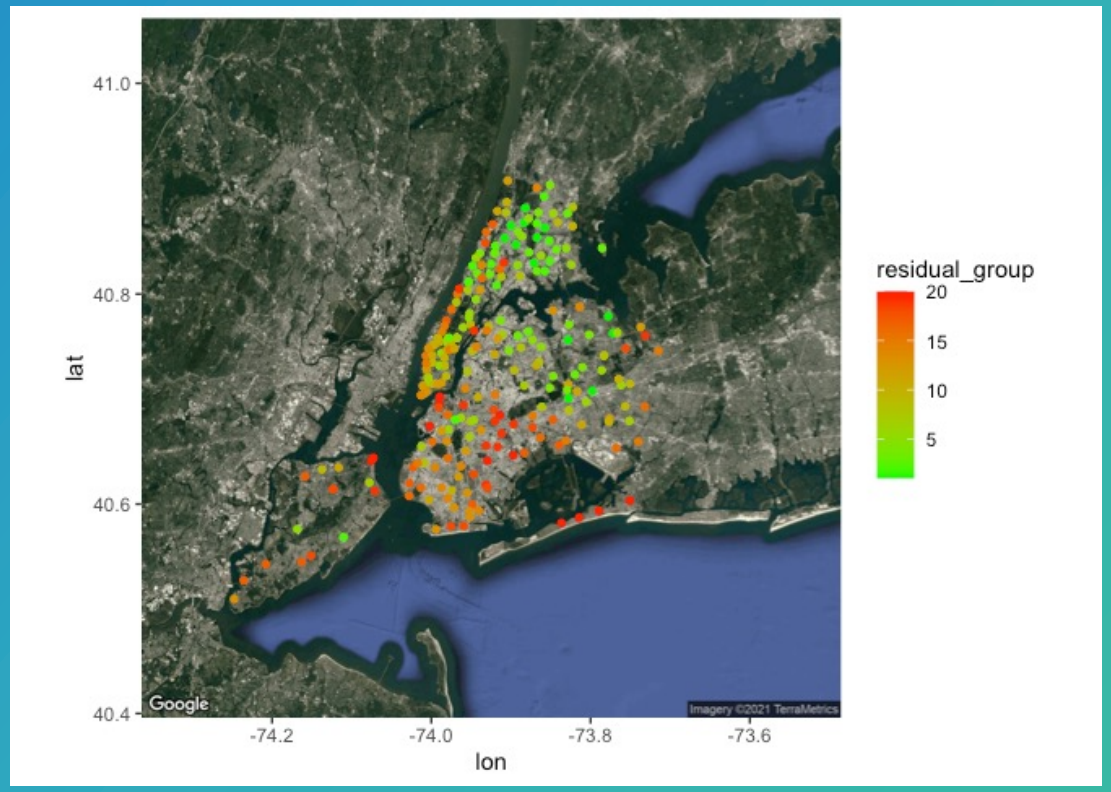
- Final model:
  - No log transformation
  - Covariates (borough and zip)
  - 2 polynomial
  - **.83 Adjusted R2**
  - **294 RSME (<5 Minutes)**





ANALYZING RESIDUALS

- Residuals are (roughly) normally distributed about zero as expected
- Increase in residuals in the middle values (better accuracy for shorter or longer trips)
- Borough paths provide some insight:
  - Routes in same borough have smaller variance
  - In general, borough groups are clustered together, as expected

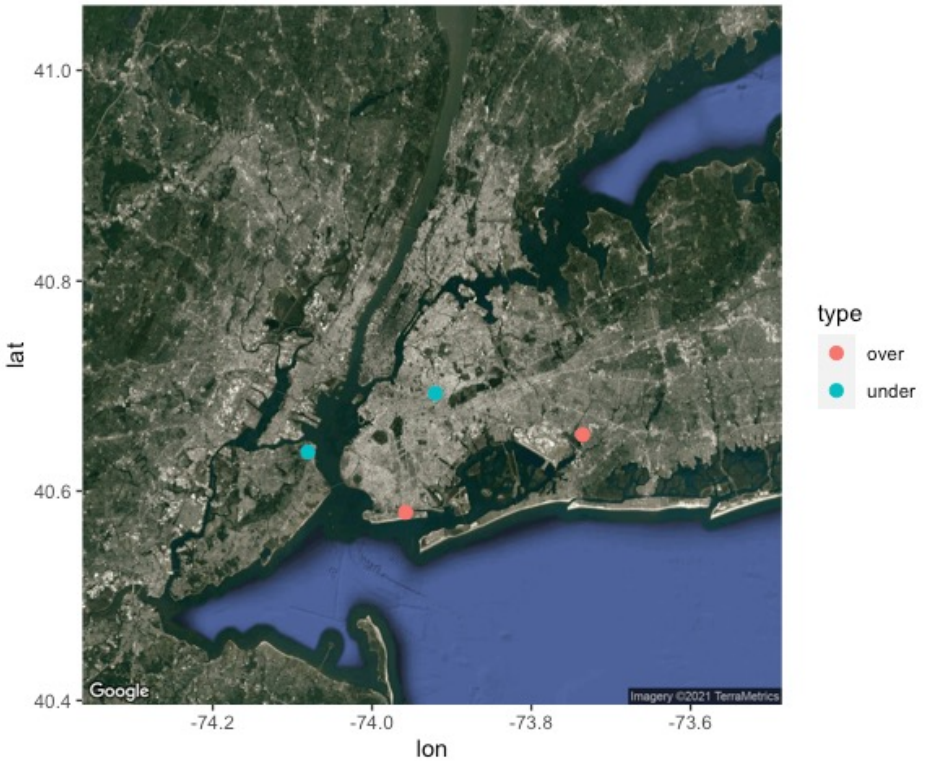


ANALYZING RESIDUALS

- Areas with the largest estimation errors include Manhattan, Staten Island, and (inner) Brooklyn

- Likely reasons for poor performance due to:
  - High population density areas
  - “Land locked” areas
  - Single Bridge areas
  - Large obstruction areas

ANALYZING RESIDUALS



- Greatest under-estimate (23 minutes)
  - Despite the short straight line distance, these points are separated by a river.
- Greatest over-estimate (16 minutes)
  - Longer straight line distance influences the model, but there are no obstacles in this route.

THANK
YOU