A Data Analysis of Austin's Construction Permits

Anika Shankar and Sofia Canihuante

Professor Guyot

SDS322E, Semester: Spring 2025

TEXAS
The University of Texas at Austin

Statistics
+ Data
Sciences

Introduction

- **Background:** Austin continues to experience a boom in new construction, particularly in residential sectors. This project seeks to identify permit value trends and assess if specific permit classes dominate.
- **Motivations:** As students living near downtown Austin, We are interested in analyzing construction trends in Austin to better understand the city's growth patterns.
- **Variables of Interest**:
- Outcome variable: valuation
  - number = permit_number,
  - class = permit_class,
  - location = permit_location,
  - units = housing_units,
  - type = permit_type_desc,
  - zipcode = original_zip
- **Prediction:** Expect single-family houses to dominate the residential construction category, predict a positively skewed distribution for job valuation due to outliers in high-end developments.
- **Research Questions:**
  - RQ 1: What is the distribution of total job valuation and permit class for new residential building permits issued in Austin during the 2022 calendar year?
  - RQ 2: Do most residential construction permits in Austin fall above or below a $1M valuation? And are there any outliers in terms of location?

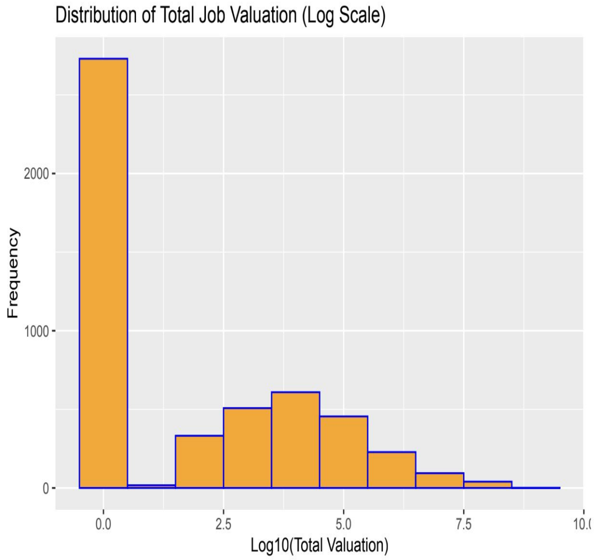
Methods

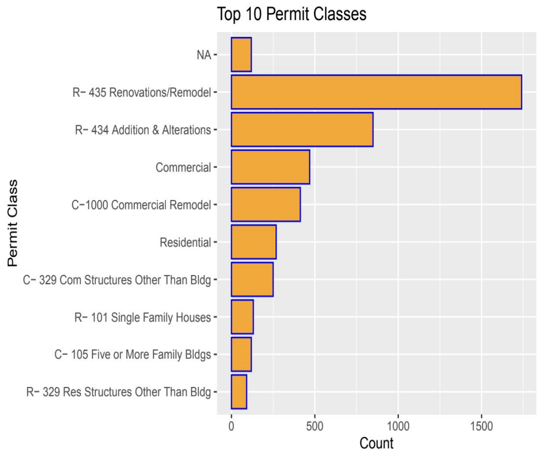
- We loaded the dataset from the **City of Austin Open Data portal** and named it permits.
- From there we modified the dataset into **permits_wrangled** from permits with the variables below using the **select() and filter()**.
    - number(from permits_number), class(from permits_class), location(from permits location), valuation(from total_job_valuation), units, zipcode(from original_zip), and excluded valuation values of null and less than 0 with !is.na() and >.
- From this wrangling that we performed using nrow() on both datasets to store the number of rows as new data and comparing them using (;) we started out with 50,000 observations and ended with 4842 observations.
- Model: Linear Regression was used to make predictions(more on this later).

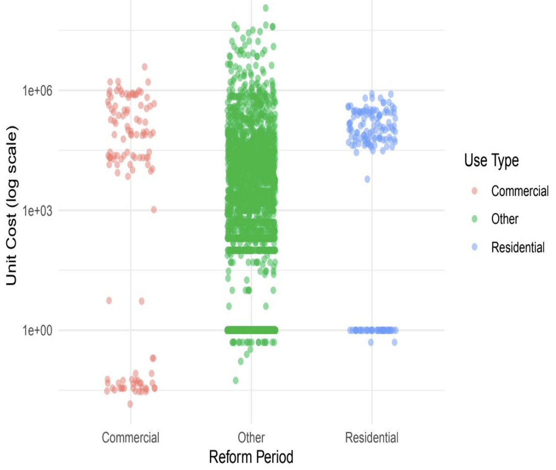
Results RQ 1:

Summary Statistics
Distribution of Job Valuation($):

```
##       Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##          1         1         1    897633      7000 540000000
```







- The bar plot of the job valuation distribution shows that most frequent valuation cost is on the lower end but with a center around 4 log10 cost units.
- The bar plot of the permit class distribution shows that the most frequent class is the R-435 Renovations/Remodeling class with close to 2000 observations.
- Multivariate Graph 1:
  - The job valuation unit cost had the most amount of range for the commercial family building classes while the residential family buildings had less range of data and higher values but overall the other classes besides those two had the most amount of data centered in 1e+05 . Commercial classes had a higher upper range of values than Residential class overall with more values in log unit cost of 1e+06 and 1e+04.

Model RQ 1: Linear Regression

```r
# Wrangle the data to create reform_period
filtered_data <- permits_wrangled %>%
  filter(
    reform_period = case_when(
      class == "R- 101 Single Family Houses" & class == "R-103 Two Family Bldgs" ~ "Residential",
      class == "C- 101 Single Family Houses" & class == "C-105 Five or More Family Bldgs" ~ "Commercial"
    )
  )
class_lm <- lm(valuation ~ class + reform_period, data = filtered_data)
summary(class_lm)
#Represent the relationship between valuation and reform period
filtered_data |>
  ggplot(aes(x = reform_period, y = valuation)) +
  # Consider a linear regression model
  geom_smooth(method = "lm", se = FALSE, size = 2)
  geom_point(size = 4) +
  labs(x = "Residential Class Type",
       y = "Job Valuation",
       title = "Linear regression model to predict Job Valaution based on Residential permit classes")
```


```
Call:
lm(formula = valuation ~ class + reform_period, data = filtered_data)

Residuals:
     Min       1Q   Median       3Q       Max
-14179559   -56234   -22034     1592 125738441

Coefficients: (2 not defined because of singularities)
                                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  162910     127043   1.282 0.199764
classC- 103 Two Family Bldgs                 142984     483274   0.296 0.767341
classC- 104 Three & Four Family Bldgs       2265998     233753   9.694  < 2e-16 ***
classC- 105 Five or More Family Bldgs       1724044     204467   8.432  < 2e-16 ***
classC- 106 Mixed Use                      14098649     497134  28.360  < 2e-16 ***
classC- 213 Hotels, Motels, & Tourist Cabins 3811461    737663   5.167 2.43e-07 ***
classC- 214 Other Nonhousekeeping Shelter   5046370     869108   5.806 6.61e-09 ***
classC- 318 Amusement, Social & Rec Bldgs    197950     438340   0.452 0.651576
classC- 319 Churches and Othr Religious Bldgs 1192996   497134   2.400 0.016428 *
classC- 320 Industrial Bldgs                 430111     593417   0.725 0.468592
classC- 321 Pkg Garage Bldg & Open Deck     1881899     373285   5.041 4.71e-07 ***
classC- 322 Service Station & Repair Garage  828101     653308   1.268 0.204993
classC- 323 Hospital & Institutional Bldgs 14437090    1926703   7.493 7.39e-14 ***
classC- 324 Office, Bank & Professional Bldgs 1759485   214352   8.208 2.57e-16 ***
classC- 325 Public Works & Utilities Bldgs  2163784     691481   3.129 0.001759 **
classC- 326 Schools & Other Educational Bldgs 846453    319522   2.649 0.008085 **
classC- 327 Stores & Customer Services       816339     231244   3.530 0.000417 ***
classC- 328 Commercial Other Nonresident Bldg 21955     185089   0.119 0.905579
classC- 329 Com Structures Other Than Bldg   -53808     199581  -0.270 0.787471
classC- 437 Addn, Alter, Convn-NonRes       1191646     237306   5.022 5.23e-07 ***
classC- 647 Demolition 3 and 4 Family Bldgs -152077    1117208  -0.136 0.891728
classC- 649 Demolition All Other Bldgs Com  -124726     211118  -0.591 0.554681
classC-1000 Commercial Remodel              -103496     145308  -0.712 0.476328
classC-1001 Commercial Finish Out           -120204     353337  -0.340 0.733717
classC-2000 Relocation Commercial           -157910    1926703  -0.082 0.934681
classCom. Driveway                          1822224    1365343   1.335 0.182034
classCom. Driveway & Sidewalk               2479977     293047   8.463  < 2e-16 ***
classCom. Driveway,Sidewalk,Curb,Gutter     3013712     737663   4.085 4.44e-05 ***
classCom. Sidewalk                          6776350    1365343   4.963 7.07e-07 ***
classCom. Sidewalk, Curb, Gutter            4667397    1365343   3.418 0.000633 ***
classCommercial                             -154627     161729  -0.956 0.339056
classR- 101 Single Family Houses             -21468     132028  -0.163 0.870837
classR- 102 Secondary Apartment              -53033     333032  -0.159 0.873482
classR- 103 Two Family Bldgs                  70397     209099   0.337 0.736375
classR- 329 Res Structures Other Than Bldg  -112004     157531  -0.711 0.477105
classR- 330  Accessory Use to Primary       -127202     209594  -0.607 0.543937
classR- 434 Addition & Alterations          -105876     139333  -0.760 0.447350
classR- 435 Renovations/Remodel             -152422     189486  -0.804 0.421191
classR- 436 Addn to increase housing units    -1514     737663  -0.002 0.998363
classR- 437 Residential Boat Dock           -144160    1365343  -0.106 0.915914
classR- 438 Residential Garage/Carport Addn -135850     205798  -0.660 0.509198
classR- 645 Demolition One Family Homes     -156015     167178  -0.933 0.350729
classR- 646 Demolition Two Family Bldgs     -154252     353337  -0.437 0.662444
classR- 649 Demolition All Other Bldgs Res  -156093     213795  -0.730 0.465347
classR-2001 Relocation Residential          -151497     264018  -0.574 0.566110
classRes. Driveway                           -46559     969614  -0.048 0.961703
classRes. Driveway & Sidewalk                -61575     497134  -0.124 0.901429
classRes. Driveway, Sidewalk, Curb, Gutter   187090    1926703   0.097 0.922646
classRes. Sidewalk                           457090    1365343   0.335 0.737799
classResidential                            -161420     169319  -0.953 0.340441
classSign Permit                            -159857     483274  -0.331 0.740819
reform_periodOther                               NA         NA      NA       NA
reform_periodResidential                         NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1923000 on 8559 degrees of freedom
  (92 observations deleted due to missingness)
Multiple R-squared:  0.1556,	Adjusted R-squared:  0.1507
F-statistic: 31.55 on 50 and 8559 DF,  p-value: < 2.2e-16
```

The model explains 15.5% of observation job valuation cost

($R^2 = 0.1556$), indicating weak predictive power for a permit

class and reform period model.

Multivariate Graph:

Based on the visualization there is no linear relationship between

reform_period and job valuation validating the summary results.

Significant Predictors of ($p<0.05$): All in the commercial class with family buildings and communal

structures being most significant of ($p<2e-16$). Reform_Period had undefined effect as shown by null values..

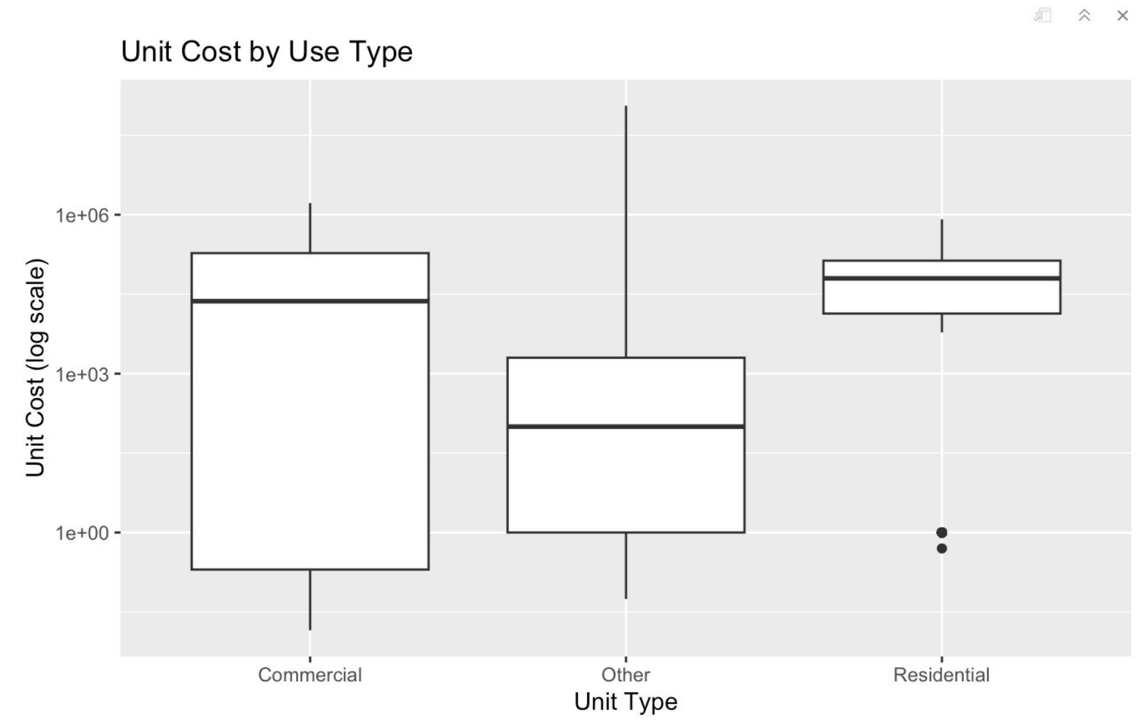
Cross Validation RQ 1:

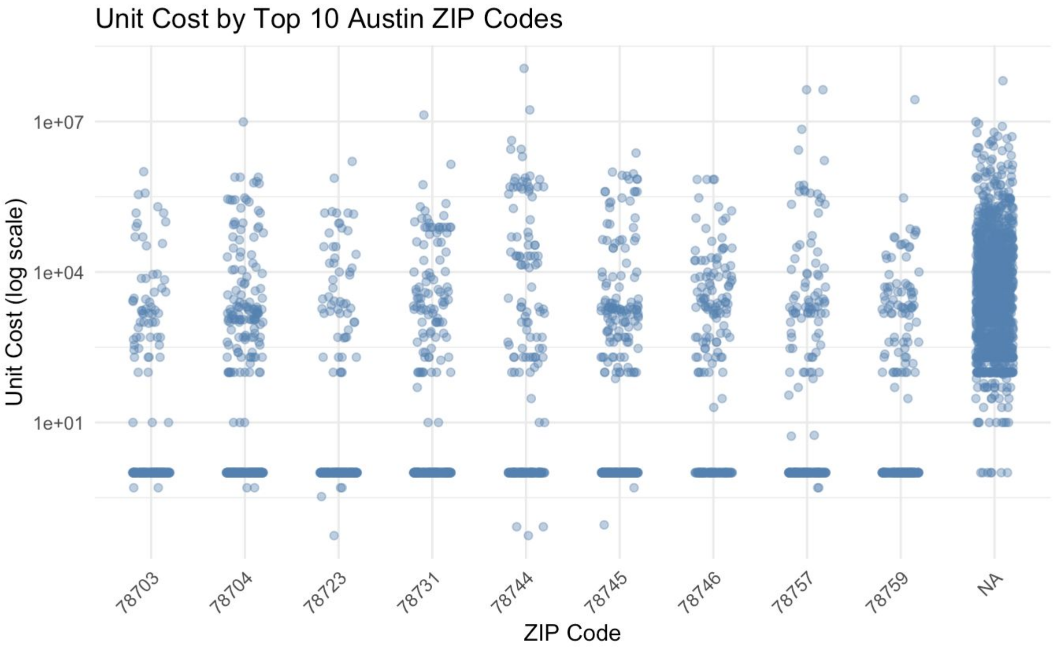
```r
set.seed(123)
# 1. prepare cleaned up data
model_set <- filtered_data |>
  dplyr::select(valuation, class, reform_period) |>
  na.omit()
#2. create the k folds
folds <- createFolds(model_set$valuation, k = 5, list = TRUE)
#3.Store results
results <- data.frame(
  Fold = 1.5,
  RMSE = numeric(5),
  R_Squared = numeric(5),
  MAE = numeric(5)
)
#4 Performing Cross Validation
for(i in 1:5){
  #split data into train and test model
  train_data <- model_set[-folds[[i]], ]
  test_data <- model_set[folds[[i]], ]

  cross_model <- lm(valuation ~ class + reform_period, data = filtered_

  #Predictions
  predicter <- predict(cross_model, newdata = test_data)
  real <- test_data$valuation

  #Cross Validation stats
  results$RMSE[i] <- sqrt(mean((real - predicter)^2))
  results$R_Squared[i] <- cor(real, predicter)^2
  results$MAE[i] <- mean(abs(real - predicter))
}
#Summary Statistics

summarize_cross <- results |>
  summarise(
    Mean_RMSE = mean(RMSE),
    SD_RMSE = sd(RMSE),
    Mean_R2 = mean(R_Squared),
    SD_R2 = sd(R_Squared),
    Mean_MAE = mean(MAE)
  )
print(results)
cat("\nSummary Statistics:\n")
print(summarize_cross)
```

```
> print(results)
  Fold      RMSE  R_Squared      MAE
1  1.5 3624435.7 0.21025694 345235.9
2  1.5 1116970.9 0.18712002 261088.6
3  1.5 1175937.1 0.08723198 221413.0
4  1.5 1273028.7 0.19593621 263879.1
5  1.5  994984.1 0.23942689 240014.7
> cat("\nSummary Statistics:\n")

Summary Statistics:
> print(summarize_cross)
  Mean_RMSE SD_RMSE   Mean_R2      SD_R2 Mean_MAE
1   1637071 1115523 0.1839944 0.05761472 266326.2
```

- The five-fold cross-validation demonstrates unstable model performance with an average R^2 of 0.18 with range of 0.12-0.24 and consistent root mean squared error 1637071$.
- This lack of consistency across the data subsets shown suggests the model does not generalize with new data points in the Austin permits data.

Results - RQ 2



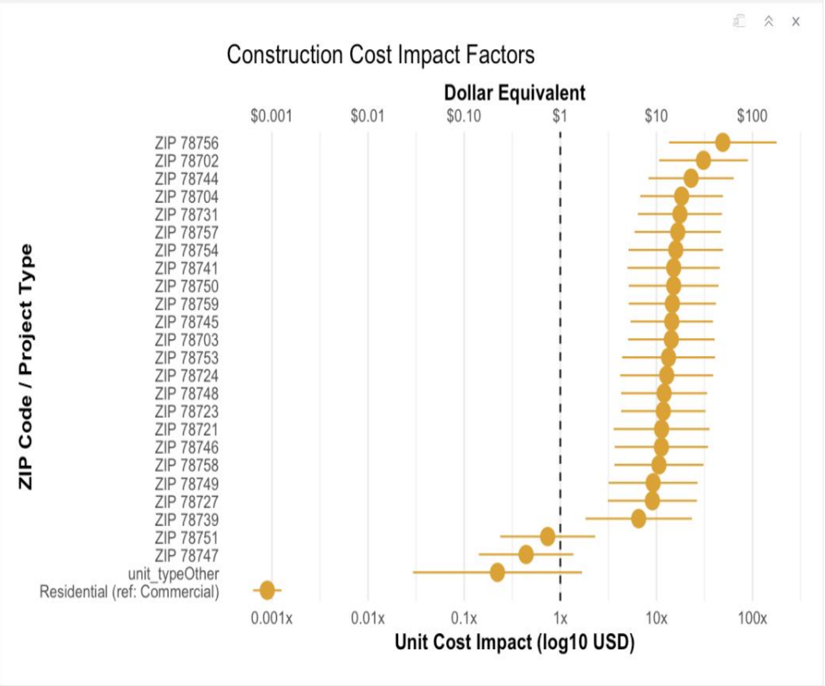


Summary Statistics: Residential projects dominated (80% of permits) but commercial drove cost extremes ($540M outliers). Log-scale visuals revealed 78704/78702 as high-variation zones with 10-100× cost spreads.

Multivariate: Commercial vs. Residential (Unit cost by Unit Type): The boxplots revealed commercial unit costs 10–100× higher than residential. The model quantified this: Commercial projects cost 1,120× more per unit ($p < 0.001$). Residential costs clustered below $100K/unit (per the log-scale histograms).

Multivariate: ZIP Code Hotspots (Unit Cost by Zip Code): The jitter plots highlighted 78704/78702/78756 as high-cost outliers. Regression (next slide) confirmed: 78756 (Tech Corridor): 49× baseline cost (*p* < 0.001). 78747/78751: No significant premium (aligned with EDA's low-cost clusters).

Model RQ 2: Linear Regression




```r
# Clean and transform data
model_data <- permits %>%
  filter(!is.na(total_job_valuation)) %>%
  filter(total_job_valuation > 0) %>%
  filter(!is.na(housing_units)) %>%
  filter(housing_units > 0) %>%
  mutate(
    unit_cost = total_job_valuation / housing_units,
    log_unit_cost = log10(unit_cost),
    unit_type = case_when(
      str_detect(permit_class, "R-") ~ "Residential",
      str_detect(permit_class, "C-") ~ "Commercial",
      TRUE ~ "Other"
    ),
    zip_code = as.factor(as.character(original_zip))
  ) %>%
  # Focus on top ZIPs
  group_by(zip_code) %>%
  filter(n() >= 50) %>%  # Only keep ZIPs with at least 50 permits
  ungroup()

# Fit linear model
cost_model <- lm(log_unit_cost ~ unit_type + zip_code, data = model_data)

# Model summary
summary(cost_model)
```

```
Call:
lm(formula = log_unit_cost ~ unit_type + zip_code, data = model_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.2701 -0.4187 -0.3519 -0.1624  6.1332 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)          2.24550    0.19835  11.321  < 2e-16 ***
unit_typeOther      -0.65361    0.44845  -1.457   0.1451    
unit_typeResidential -3.04902   0.07543 -40.424  < 2e-16 ***
zip_code78702        1.48975    0.23576   6.319 2.99e-10 ***
zip_code78703        1.15538    0.22956   5.033 5.09e-07 ***
zip_code78704        1.26258    0.21947   5.753 9.58e-09 ***
zip_code78721        1.05369    0.25487   4.134 3.65e-05 ***
zip_code78723        1.07165    0.22425   4.779 1.84e-06 ***
zip_code78724        1.10714    0.24691   4.484 7.58e-06 ***
zip_code78727        0.95809    0.23624   4.056 5.12e-05 ***
zip_code78731        1.24486    0.22327   5.576 2.67e-08 ***
zip_code78739        0.81556    0.28241   2.888   0.0039 ** 
zip_code78741        1.17955    0.24510   4.812 1.56e-06 ***
zip_code78744        1.36121    0.22623   6.017 1.97e-09 ***
zip_code78745        1.15995    0.21889   5.299 1.24e-07 ***
zip_code78746        1.05176    0.24747   4.250 2.20e-05 ***
zip_code78747       -0.35562    0.25089  -1.417   0.1564    
zip_code78748        1.08021    0.22832   4.731 2.33e-06 ***
zip_code78749        0.96588    0.23595   4.094 4.35e-05 ***
zip_code78750        1.17951    0.23815   4.953 7.68e-07 ***
zip_code78751       -0.13106    0.25189  -0.520   0.6029    
zip_code78753        1.12619    0.24664   4.566 5.15e-06 ***
zip_code78754        1.20172    0.25083   4.791 1.73e-06 ***
zip_code78756        1.69063    0.28571   5.917 3.61e-09 ***
zip_code78757        1.22222    0.22925   5.331 1.04e-07 ***
zip_code78758        1.02678    0.23623   4.347 1.43e-05 ***
zip_code78759        1.16626    0.23115   5.046 4.77e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.464 on 3272 degrees of freedom
  (1456 observations deleted due to missingness)
Multiple R-squared:  0.3636,	Adjusted R-squared:  0.3586 
F-statistic: 71.91 on 26 and 3272 DF,  p-value: < 2.2e-16
```

The model shows commercial construction projects cost approximately 1,120 times more per unit than residential projects ($p < 0.001$).

Geographic analysis reveals significant location-based variations, with the tech corridor (78756) showing the strongest cost premium at 49 times baseline values, followed by downtown ZIP codes 78704 (18×) and 78702 (31×).

The model explains 36% of observed cost variation ($R^2 = 0.36$), indicating moderate predictive power for a socioeconomic model.

Cross Validation RQ 2:

```
{r}
set.seed(123) # For reproducibility

# 1. Prepare the data (using your cleaned top_zips dataset)
model_data <- top_zips %>%
  select(log_unit_cost, unit_type, zip_code) %>%
  na.omit() # Remove any remaining NAs

# 2. Create stratified folds (maintains class balance)
folds <- createFolds(model_data$log_unit_cost, k = 5, list = TRUE)

# 3. Initialize results storage
results <- data.frame(
  Fold = 1:5,
  RMSE = numeric(5),
  R_squared = numeric(5),
  MAE = numeric(5) # Adding Mean Absolute Error
)

# 4. Run cross-validation
for(i in 1:5) {
  # Split data
  train_data <- model_data[-folds[[i]], ]
  test_data <- model_data[folds[[i]], ]

  # Train model
  cv_model <- lm(log_unit_cost ~ unit_type + zip_code, data = train_data)

  # Make predictions
  predictions <- predict(cv_model, newdata = test_data)
  actual <- test_data$log_unit_cost

  # Calculate metrics
  results$RMSE[i] <- sqrt(mean((actual - predictions)^2))
  results$R_squared[i] <- cor(actual, predictions)^2
  results$MAE[i] <- mean(abs(actual - predictions))
}

# 5. Compute summary statistics
summary_stats <- results %>%
  summarise(
    Mean_RMSE = mean(RMSE),
    SD_RMSE = sd(RMSE),
    Mean_R2 = mean(R_squared),
    SD_R2 = sd(R_squared),
    Mean_MAE = mean(MAE)
  )

# 6. Print results
print(results)
cat("\nSummary Statistics:\n")
print(summary_stats)
```

- Five-fold cross-validation demonstrates stable model performance with an average R^2 of 0.35 (range: 0.29-0.40) and consistent root mean squared error (1.48 ± 0.06 log10 units).

- This consistency across data subsets suggests the model generalizes well to new observations within Austin's permit data framework, though predictive accuracy remains limited for high-value commercial projects.

Mean_RMSE <dbl>	SD_RMSE <dbl>	Mean_R2 <dbl>	SD_R2 <dbl>	Mean_MAE <dbl>
1.477149	0.06426713	0.3508782	0.05304796	0.8924462

Discussion RQ1:

Key Findings:
- The question, “What is the distribution of total job valuation and permit class for new residential building permits issued in Austin during the 2022 calendar year?” was answered in this analysis.
- The EDA multivariate scatterplot shows that their is small difference in the reform period class as well as a lack of variation in the other classes which was confirmed by the to the linear regression.
- The model explains 15.5% of the cost valuation variation with prediction errors of +/-$1923000 dollars shows that is may work for small clusters of projects than doing city wide precise projections

Surprises and Limitations:
- While commercial buildings overall seem to be vary more as expected by the 2022 reforms which focused on affordable housing it didn't seem to make a big overall difference in cost valuation
- While their were more commercial classes the most dominant class observation count was R-435 Renovations/Remodeling with close to 2000 observations.
- Reform_Period did not even register to the model with the NA values which is a limitation of the model as we only included 4 values in that category focusing on family buildings instead of a more holistic model.(e.g. R-101 or C-101 Single Family Houses)

Recommendations to the City:
- Based on these findings my recommendations to the city would be to expand their residential class buildings if they value diversity in building projects and no impact in job valuation cost. If they value impacts on the cost valuation for the overall projects they should include more commercial class buildings as most of the significance of the model (p-value<0.05) was shown in the commercial classes. More family buildings should be built in order to gain actionable insights as the R-101 Single Family house was low on the bar chart for observations at only less than 250 observations.

Discussion continued - RQ 2

1. Key Findings and the relation to RQ 2

- The question, “Do most residential construction permits in Austin fall above or below a $1M valuation? And are there any outliers in terms of location?” was answered in this analysis…
- The analysis proves commercial construction costs are 1,120× higher per unit than residential (*p* < 0.001). The EDA’s jitter plots first revealed this divide, showing 78756 (Tech Corridor) as the most extreme outlier at 49× baseline costs—later confirmed by the linear regression model.
- While the model explains 36% of cost variation, prediction errors of ±$30/unit mean it’s better for neighborhood-level estimates than precise projections.

2. Surprises & Limitations

- The expectation of residential projects dominating was accurate, but the $540M commercial outliers (seen in the histograms) were shocking.
- Its surprising that 78747 and 78751 resist the cost spikes seen in nearby downtown ZIPs.
- The analysis was limited by inconsistent permit class labels (e.g., "R-101" vs. "R101")—standardizing these would improve future work.

3. Recommendations for the City

Based on these findings, the recommendation to the city would be: Prioritize affordable housing in 78747/78751, because of tight cluster of sub-$100K residential units and lack of cost premium make them prime for affordable housing. In addition, monitor commercial overdevelopment in 78756/78704, as extreme commercial outliers (31-49 × baseline costs) risk pricing out affordable development if unregulated. My ZIP code jitter plots and regression results both support this.

Reflection

Cleaning the permits data posed significant challenges, particularly standardizing inconsistent permit classes and handling missing unit counts. Incorporating ZIP codes required manual API adjustments but enabled critical spatial analysis. Log transformations proved essential for visualizing extreme cost outliers. The project strengthened our data wrangling and visualization skills while providing actionable insights into Austin's development patterns.

Acknowledgements

Anika led data cleaning, ZIP code integration, and initial analysis. Anika focused on the information for RQ 2, while Sofia focused on RQ 1. Sofia refined visualizations and report formatting. Both members collaborated on interpretation and debugging. We'd like to thank our TAs and Professor Guyot for their support, as well as the City of Austin.

References

- City of Austin, Texas - data.austintexas.gov. "Issued Construction Permits: Open Data: City of Austin Texas." *Data.AustinTexas.Gov - The Official City of Austin Open Data Portal*, 28 Apr. 2025, data.austintexas.gov/Building-and-Development/Issued-Construction-Permits/3syk-w9eu/about_data. Accessed 20 Apr. 2025.
- Lee, Jonathan (Ph. D. in community and regional planning). "Public Land for Public Good: Leveraging City-Owned Property to Create Affordable Housing in Austin, Texas." TSW, 1 Aug. 2023, repositories.lib.utexas.edu/items/b10e880a-3876-4323-8015-e59314c94c0f. Accessed 28 Apr. 2025.
- Swiatecki, Chad. "As Construction Slowdown Continues, a New Workforce Housing Project Takes Shape." *Austin Monitor*, 2 Aug. 2024, www.austinmonitor.com/stories/2024/08/as-construction-slowdown-continues-a-new-workforce-housing-project-takes-shape/. Accessed 20 Apr. 2025.
- www.nytimes.com/2012/08/15/realestate/commercial/in-downtown-austin-a-wave-of-construction.html. Accessed 20 Apr. 2025.