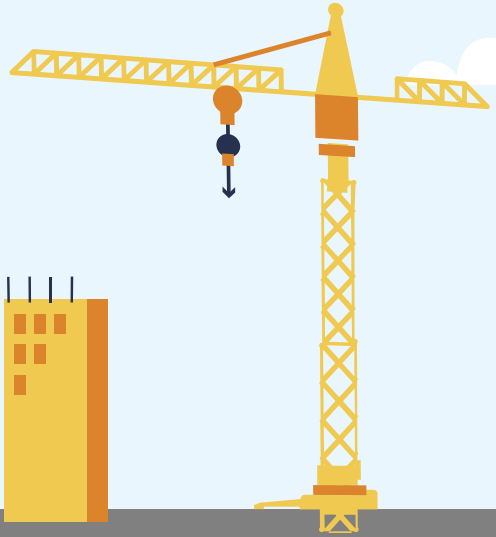
ADEC 7430

Building Permit Predictions

Alice Tang



Presentation Outline

01

Introduction

02

Data Cleaning

An explanation of the decision made in this process.

03

Examining Data

04

Building the Model

Introducing the variables used in the model.

05

Transforming Variables

Expanding on the processes behind building this model.

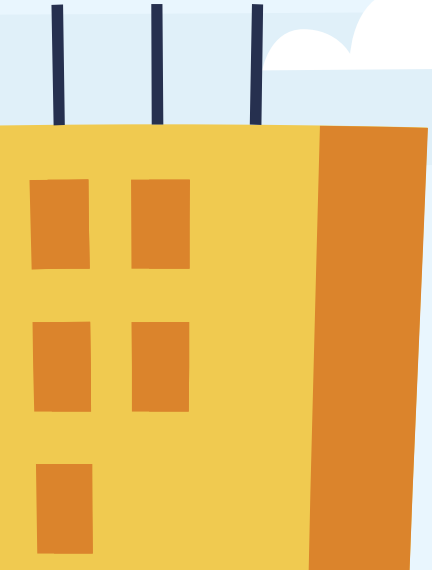
06

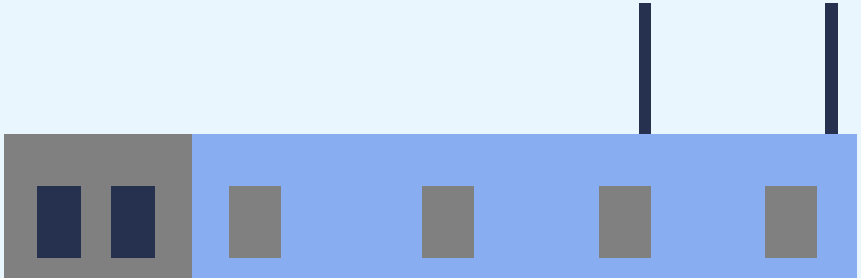
Concluding Results

Summarizing the results and reflecting on what could have been done better.

01.

Introduction



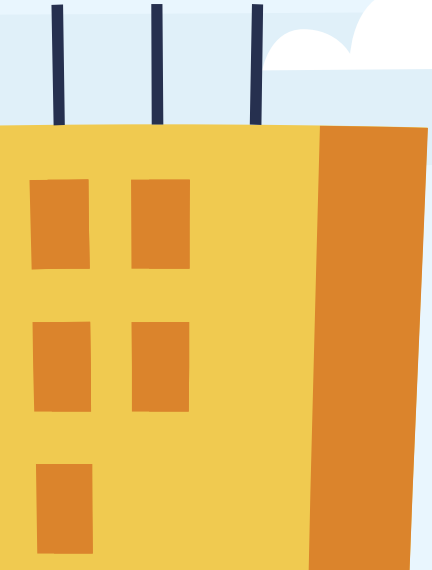


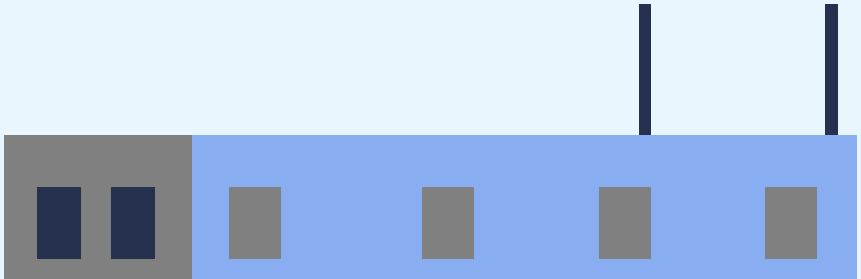
GOAL:

To build a model that would estimate how long permit finalizations would take for Chicago buildings based on characteristics in the dataset.

02.

Data Cleaning





Steps and Decisions Made in Cleaning Data

Drop Columns

- Less overwhelming

Drop Blanks

- Many “ “ values

Drop N/A Values

- Removed → large dataset
  - Less impact than if small

Convert Faulty Data Types

- Permit / Review type
  - Chr → factor w/ levels

Correct Date format

- %m/%d/%y
  - consistency

Drop Skew Values

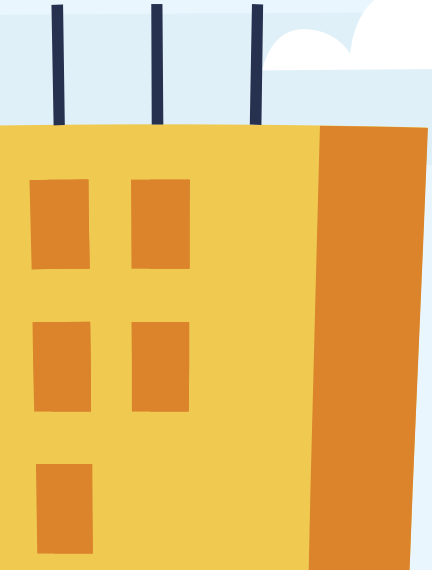
- Examined summary, min and max
  - (-) processing time and total fee
    - Most likely misentry

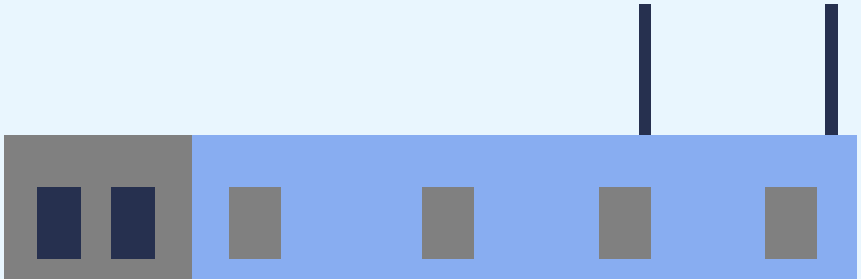
Skew Values

PROCESSING_TIME	ISSUE_DATE	TOTAL_FEE
Min. :-2876.00	Min. :2006-01-03	Min. : -11527
1st Qu.: 0.00	1st Qu.:2010-02-08	1st Qu.: 75
Median : 0.00	Median :2014-08-21	Median : 225
Mean : 22.63	Mean :2014-05-22	Mean : 935
3rd Qu.: 8.00	3rd Qu.:2018-07-16	3rd Qu.: 500
Max. : 5699.00	Max. :2022-09-27	Max. :5772092

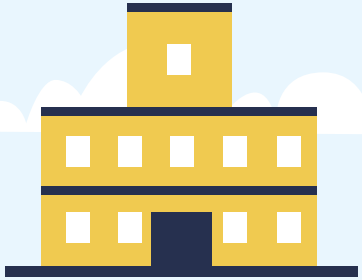
03.

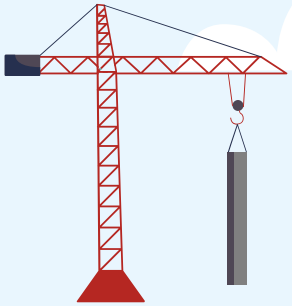
Examining Data





Number of Unique Permit and Review Types





11 Unique Permit Types

```
> length(unique(building$PERMIT_TYPE))
[1] 11
```

11 Unique Review Types

```
> length(unique(building$REVIEW_TYPE))
[1] 11
```

Distribution of Records per Permit

```
> #distribution of records per permit, will be important for determining reference group later on
> #we can see that the most were electric wiring
> table(building$PERMIT_TYPE)

  PERMIT - EASY PERMIT PROCESS       PERMIT - ELECTRIC WIRING    PERMIT - ELEVATOR EQUIPMENT  PERMIT - FOR EXTENSION OF PMT      PERMIT - NEW CONSTRUCTION
                        196212                         251040                         18003                            58                          26465
   PERMIT - PORCH CONSTRUCTION PERMIT - REINSTATE REVOKED PMT PERMIT - RENOVATION/ALTERATION          PERMIT - SCAFFOLDING                  PERMIT - SIGNS
                          3096                           3813                        143840                          8574                          45694
  PERMIT - WRECKING/DEMOLITION
                         19389
>
```

- Important to **determine reference group** for linear regression
- Electric Wiring Permit had the most → **196,212**
- For Extension of Pmt → **58 entries only**
- **Much range in the distribution**

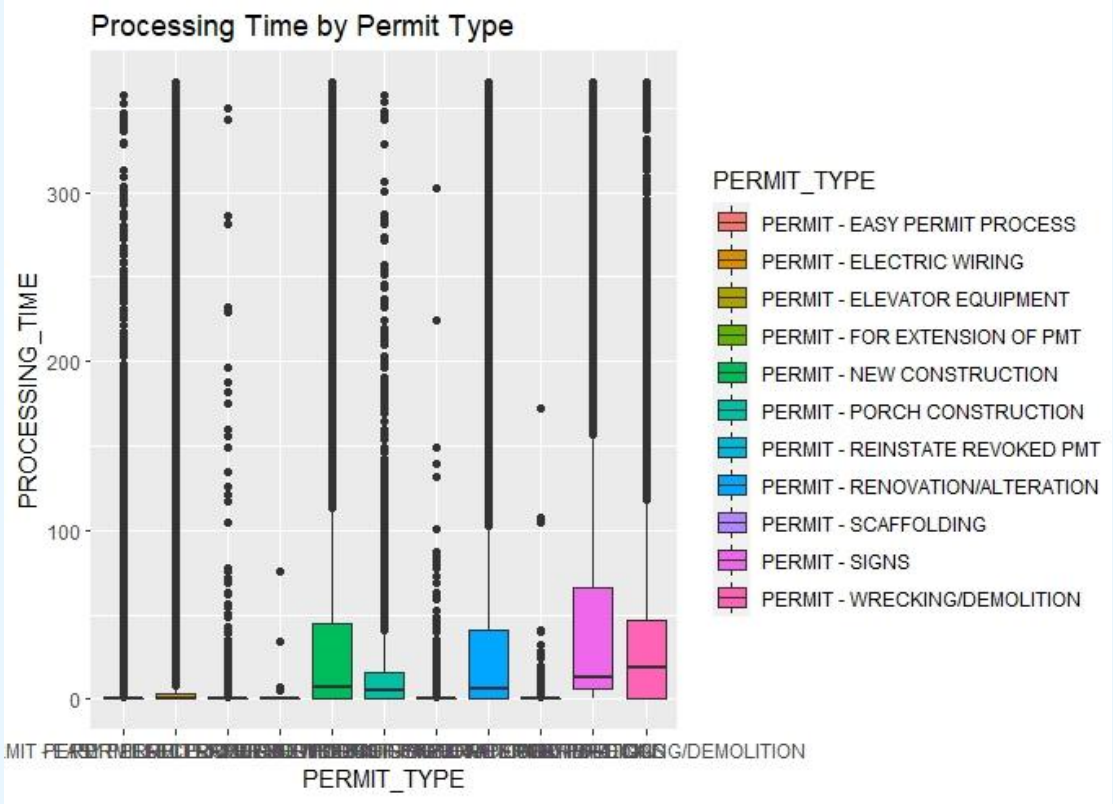
Distribution of Records per Review

```
> #distribution of records per review, will be important for determining reference group later on
> #majority of review type was easy permit web
> table(building$REVIEW_TYPE)

      CONVEYANCE DEVICE PERMIT                DEMOLITION PERMIT      DIRECT DEVELOPER SERVICES                 EASY PERMIT             EASY PERMIT WEB
                         18003                            19389                           1189                        160543                      286931
        ELECTRICAL PLAN REVIEW          FIRE PROTECTION SYSTEM                      SELF CERT                 SIGN PERMIT  STANDARD PLAN REVIEW
                          5922                             6435                          43202                        45694                      123531
TRADITIONAL DEVELOPER SERVICES
                          5345
```

- Easy Permit Web had the most → **286,931**
- Direct Developer Services → **1,189 entries only**
- **Also quite widespread → gravitates more to easy permit**
  - Easy permit → **"streamline process for small and simple building improvements"** (per the City of Chicago's official website)

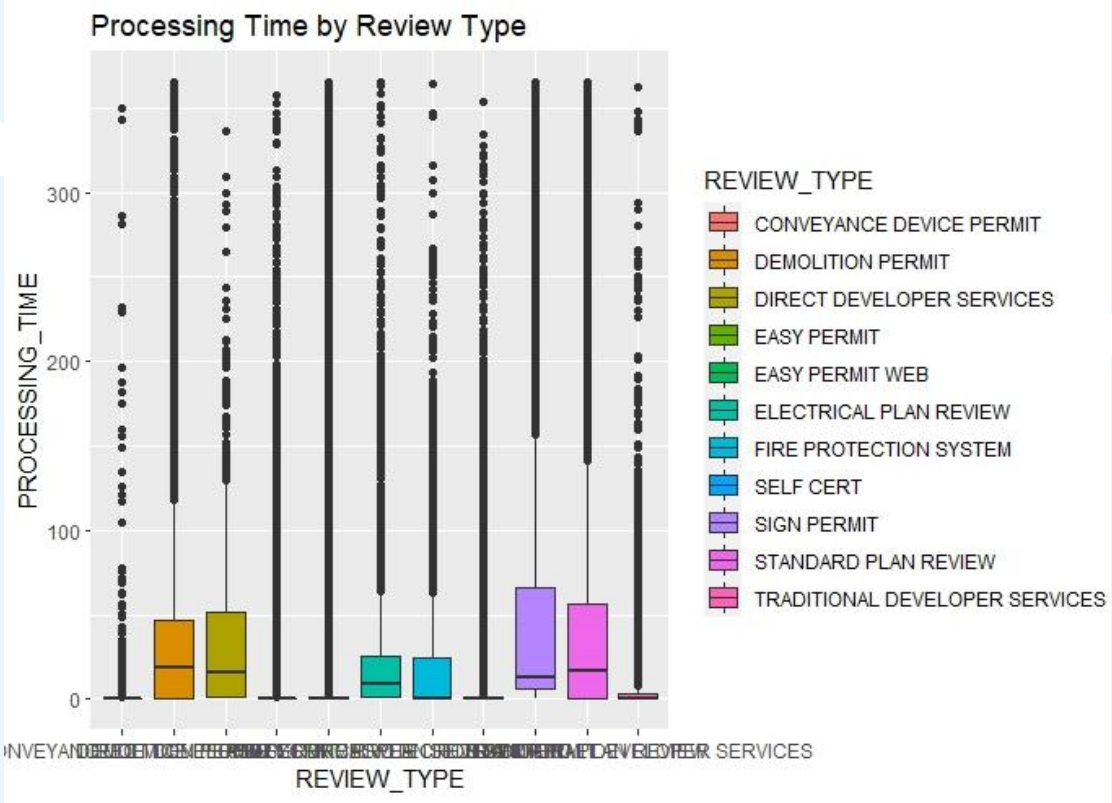
Processing Time by Type of Permit



- 1 yr was maximum
- Easy Permit Process + Electric Wiring → **skewed**
  - Many at 0 but also many entries spread across **spread across large amounts of time.**
- For some w/ visible boxes → **medians are higher/lower than others**
  - Wrecking/demolition → **higher time on avg**
- Boxes you can't see → on avg take a less amt of time but **extensive range of values**

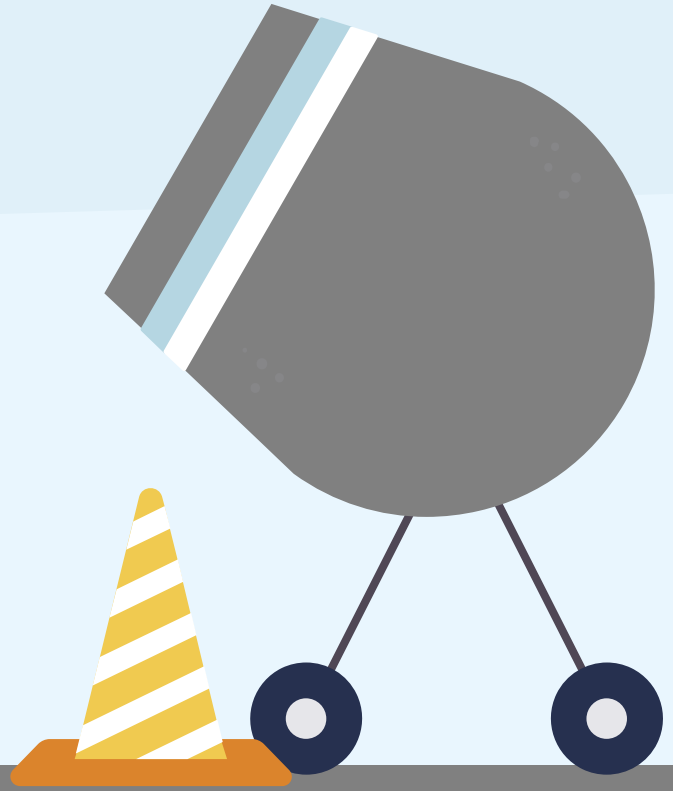
Processing Time by Type of Review

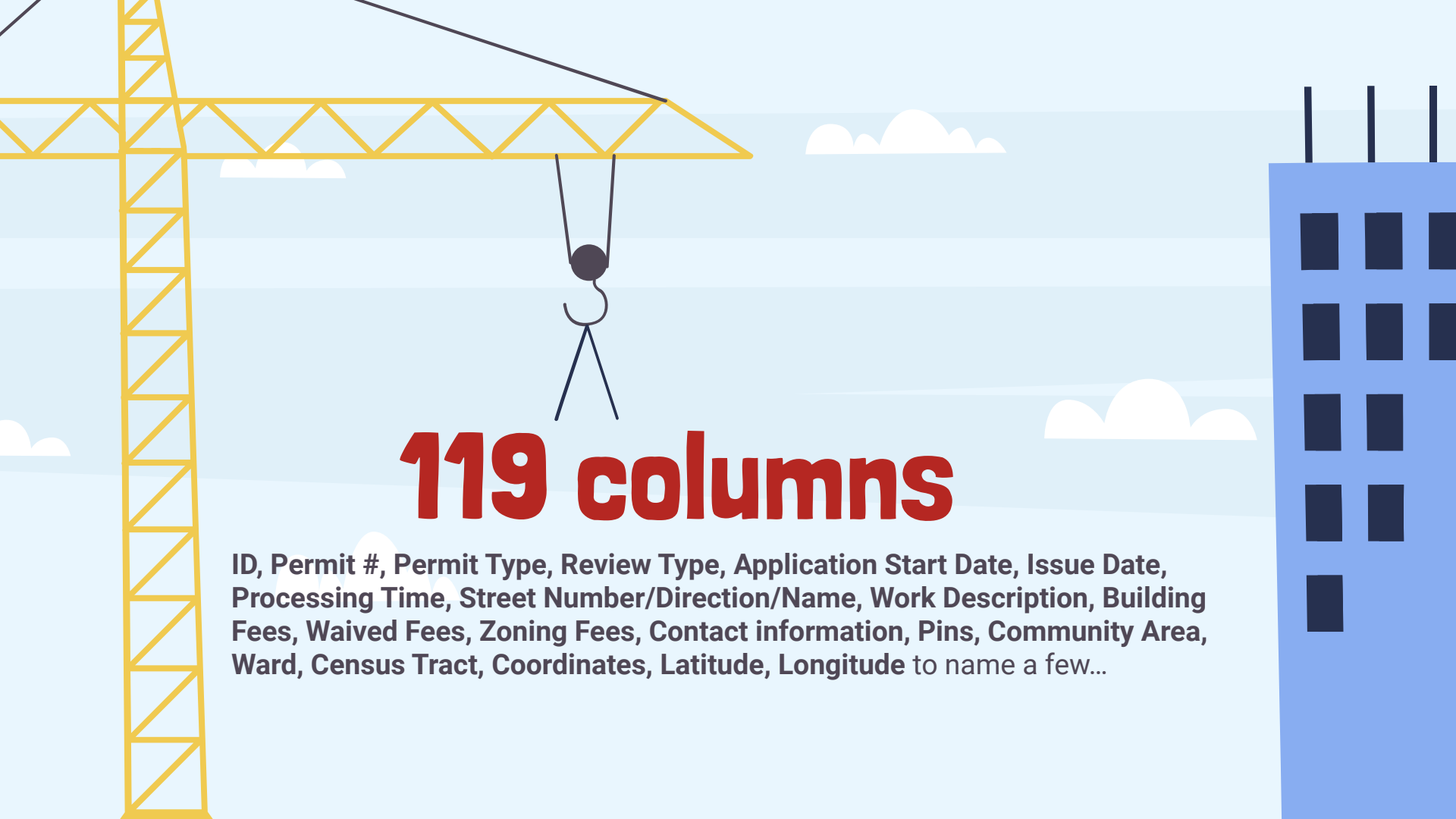
- 1 yr was maximum
- Once again, so **skewed**
  - Expected due to the **large amounts of data in certain types and less in others**
- Higher median for **Standard Plan Review** than others
  - Also interesting bc 3rd most amt of data in this type, overall takes longer
- Boxes you can't see → on avg take a less amt of time but **extensive range of values**



04.

Building the Model



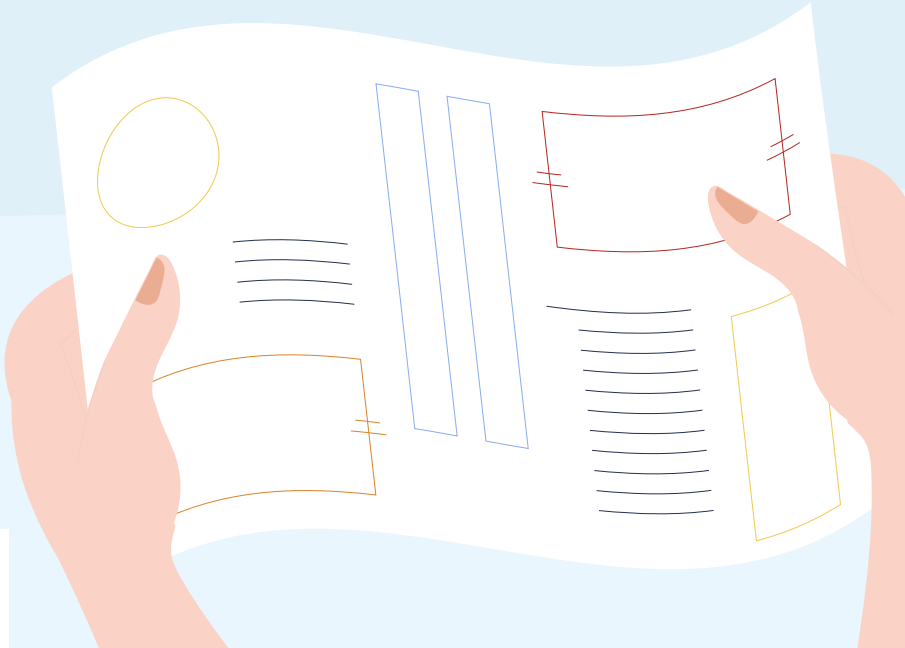


119 columns

ID, Permit #, Permit Type, Review Type, Application Start Date, Issue Date, Processing Time, Street Number/Direction/Name, Work Description, Building Fees, Waived Fees, Zoning Fees, Contact information, Pins, Community Area, Ward, Census Tract, Coordinates, Latitude, Longitude to name a few...

Variables Chosen

- Permit Type, Review Type
- Application Start Date, Processing Time (Y)
- Total Fee, Subtotal Waived

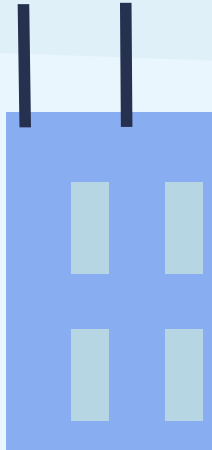


Removed: Community Area :(

```
> sum(is.na(building$COMMUNITY_AREA))
[1] 104315
>
```

05.

Transforming Variables



Grouping Decision

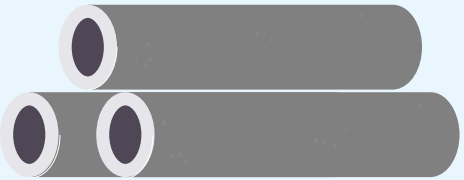
- Ran linear regression without grouping to see which ones should group tg
- Looked at p-values, **insignificant ones went into other**
- However, regardless of p-values if permits were very distinct / diff from others kept in own group.
    - Mainly **put smaller groups together** to form a larger one → **big size** → **less variance**

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.126e+03	3.425e+00	620.804	< 2e-16	***
relevel(PERMIT_TYPE, ref = "PERMIT - ELECTRIC WIRING")PERMIT - EASY PERMIT PROCESS	-1.005e+01	4.299e-01	-23.387	< 2e-16	***
relevel(PERMIT_TYPE, ref = "PERMIT - ELECTRIC WIRING")PERMIT - ELEVATOR EQUIPMENT	-9.822e+00	6.547e-01	-15.000	< 2e-16	***
relevel(PERMIT_TYPE, ref = "PERMIT - ELECTRIC WIRING")PERMIT - FOR EXTENSION OF PMT	-9.596e+00	1.058e+01	-0.907	0.364269	
relevel(PERMIT_TYPE, ref = "PERMIT - ELECTRIC WIRING")PERMIT - NEW CONSTRUCTION	-2.131e+01	7.409e+00	-2.876	0.004030	**
relevel(PERMIT_TYPE, ref = "PERMIT - ELECTRIC WIRING")PERMIT - PORCH CONSTRUCTION	-6.950e+01	7.578e+00	-9.172	< 2e-16	***
relevel(PERMIT_TYPE, ref = "PERMIT - ELECTRIC WIRING")PERMIT - REINSTATE REVOKED PMT	-8.691e+00	1.452e+00	-5.984	2.18e-09	***
relevel(PERMIT_TYPE, ref = "PERMIT - ELECTRIC WIRING")PERMIT - RENOVATION/ALTERATION	-2.443e+01	7.416e+00	-3.294	0.000989	***
relevel(PERMIT_TYPE, ref = "PERMIT - ELECTRIC WIRING")PERMIT - SCAFFOLDING	-1.156e+01	1.034e+00	-11.185	< 2e-16	***
relevel(PERMIT_TYPE, ref = "PERMIT - ELECTRIC WIRING")PERMIT - SIGNS	6.785e+01	4.326e-01	156.843	< 2e-16	***
relevel(PERMIT_TYPE, ref = "PERMIT - ELECTRIC WIRING")PERMIT - WRECKING/DEMOLITION	2.259e+01	6.324e-01	35.726	< 2e-16	***
relevel(REVIEW_TYPE, ref = "EASY PERMIT WEB")CONVEYANCE DEVICE PERMIT	NA	NA	NA	NA	
relevel(REVIEW_TYPE, ref = "EASY PERMIT WEB")DEMOLITION PERMIT	NA	NA	NA	NA	
relevel(REVIEW_TYPE, ref = "EASY PERMIT WEB")DIRECT DEVELOPER SERVICES	4.144e+01	7.853e+00	5.277	1.31e-07	***
relevel(REVIEW_TYPE, ref = "EASY PERMIT WEB")EASY PERMIT	-6.395e-03	4.517e-01	-0.014	0.988703	
relevel(REVIEW_TYPE, ref = "EASY PERMIT WEB")ELECTRICAL PLAN REVIEW	1.569e+01	1.116e+00	14.055	< 2e-16	***
relevel(REVIEW_TYPE, ref = "EASY PERMIT WEB")FIRE PROTECTION SYSTEM	6.873e+00	1.072e+00	6.410	1.45e-10	***
relevel(REVIEW_TYPE, ref = "EASY PERMIT WEB")SELF CERT	1.761e+01	7.422e+00	2.373	0.017651	*
relevel(REVIEW_TYPE, ref = "EASY PERMIT WEB")SIGN PERMIT	NA	NA	NA	NA	
relevel(REVIEW_TYPE, ref = "EASY PERMIT WEB")STANDARD PLAN REVIEW	5.581e+01	7.414e+00	7.528	5.17e-14	***
relevel(REVIEW_TYPE, ref = "EASY PERMIT WEB")TRADITIONAL DEVELOPER SERVICES	1.693e+01	7.499e+00	2.257	0.023990	*

Grouping Permits

```
#grouping the types of permits together to improve model
building$type_permit <- "OTHER"
building$type_permit[building$PERMIT_TYPE == "PERMIT - ELECTRIC WIRING"] <- "ELECTRIC WIRING"
building$type_permit[building$PERMIT_TYPE == "PERMIT - EASY PERMIT PROCESS"] <- "EASY PERMIT PROCESS"
building$type_permit[building$PERMIT_TYPE == "PERMIT - RENOVATION/ALTERATION"] <- "RENOVATION/ALTERATION"
building$type_permit[building$PERMIT_TYPE == "PERMIT - NEW CONSTRUCTION"] <- "NEW CONSTRUCTION"
building$type_permit[building$PERMIT_TYPE == "PERMIT - WRECKING/DEMOLITION"] <- "WRECKING/DEMOLITION"
building$type_permit[building$PERMIT_TYPE == "PERMIT - PORCH CONSTRUCTION"] <- "PORCH CONSTRUCTION"
building$type_permit[building$PERMIT_TYPE == "PERMIT - SIGNS"] <- "SIGNS"
building$type_permit[building$PERMIT_TYPE == "PERMIT - ELEVATOR EQUIPMENT"] <- "ELEVATOR EQUIPMENT"
```

- **Scaffolding, Reinstate Revoked PMT, For Extension of PMT grouped in “other”**



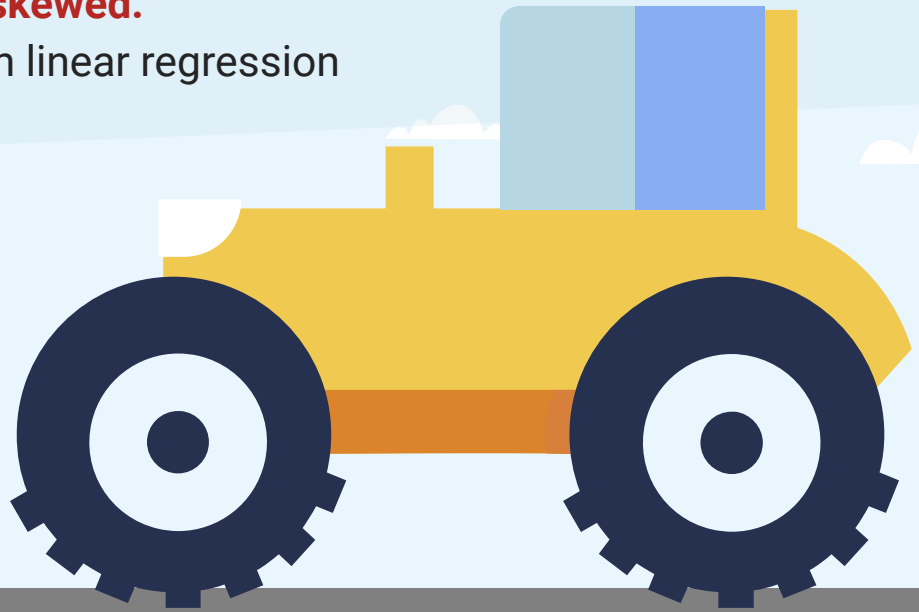
Grouping Reviews

```
#grouping the types of reviews together to improve model
table(building$REVIEW_TYPE)
building$type_review <- "OTHER"
building$type_review[building$REVIEW_TYPE == "EASY PERMIT WEB"] <- "EASY PERMIT WEB"
building$type_review[building$REVIEW_TYPE == "EASY PERMIT"] <- "EASY PERMIT"
building$type_review[building$REVIEW_TYPE == "FIRE PROTECTION SYSTEM"] <- "FIRE PROTECTION SYSTEM"
building$type_review[building$REVIEW_TYPE == "STANDARD PLAN REVIEW"] <- "STANDARD PLAN REVIEW"
building$type_review[building$REVIEW_TYPE == "SELF CERT"] <- "SELF CERT"
building$type_review[building$REVIEW_TYPE == "ELECTRICAL PLAN REVIEW"] <- "ELECTRICAL PLAN REVIEW"
building$type_review[building$REVIEW_TYPE == "TRADITIONAL DEVELOPER SERVICES"] <- "TRADITIONAL DEVELOPER SERVICES"
```

- **Sign Permit, Demolition, Conveyance Device Permit, Direct Developer Services, grouped in “other”**

Log Transformations

- Used this function: **log1p(TOTAL_FEE) + log1p(SUBTOTAL_WAIVED)**
  - As we see from our summary statistics, **TOTAL_FEE** and **SUBTOTAL_WAIVED** data is **quite skewed.**
  - Skewed data → **negative impact** on linear regression



06.

Concluding Results

Method: Using a Random Sample

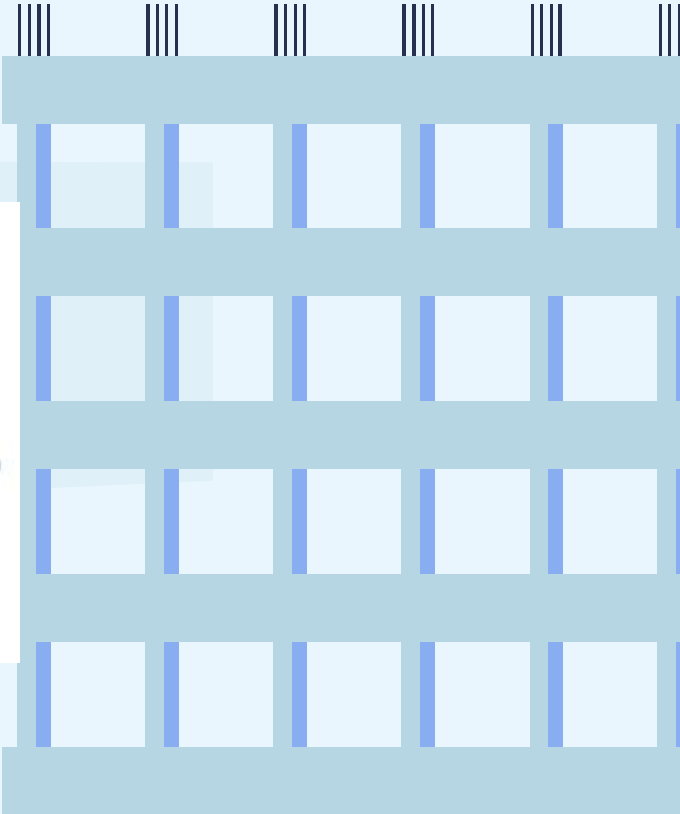
- Desire for **reproducible results**
  - **set.seed() function** good for **creating simulations** or **random objects reproduced**

```
################## Data Partition ##########################################
#will be using random sampling for this assignment.
library(caret)
#setting seed to produce a reproducible random sampling
#ask what value this should be set to
set.seed(123)

#creating training data as 70% of the dataset
random_sample <- createDataPartition(building$PROCESSING_TIME, p = 0.7, list = FALSE)

#generating training dataset from the random_sample
train <- building[random_sample, ]

#generating testing dataset from rows not included in random_sample
test <- building[-random_sample,]
```



Training and Predicting

- **Training the model for Linear Regression:**

```
model <- lm((PROCESSING_TIME) ~ relevel(factor(type_permit), ref = "ELECTRIC WIRING") +
relevel(factor(type_review), ref = "EASY PERMIT WEB")  + log1p(TOTAL_FEE) +
log1p(SUBTOTAL_WAIVED) + application_year, data = train)
```

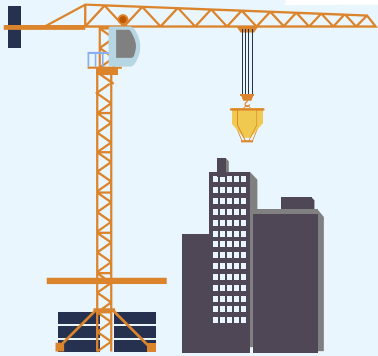
- **Predicting the target variable (processing time):**

```
predictions <- predict(model, test)
```

Model Performance Metrics

```
> #computing model performance metrics
> data.frame(R2= R2(predictions, test$PROCESSING_TIME),
+            RMSE = RMSE(predictions, test$PROCESSING_TIME),
+            MAE = MAE(predictions, test$PROCESSING_TIME))
         R2     RMSE     MAE
1 0.5241548 72.31185 24.1498
```

- **R^2**
  - How well predictor variables explain variation in response variable
  - **52.4%** of processing time are due to variation of permit type/review type, etc;
- **RMSE**
  - How well regression model predicts value of response variable
    - Lower = better model
- **MAE**
  - Average of all absolute errors
  - Lower = better model
- In general → expected since the range of outcome/predictor variables were very large



Reflecting on Potential Drawbacks/Challenges

Chosen a Different Method

- Random forest
  - Memory issues

Difficulty in Dealing with Numeric and Categorical Data

- Hard finding good predictive models to use
  - If interested in one, was only for categorical data or vice versa
    - Eg: the outlier model

Choosing Other Variables

- Other variables like community area
  - Not enough data

Cleaning the Data

- Big range of data in terms of processing time → maybe could have limited and been better at dealing with outliers

Thank you for listening!

If there are any further questions or concerns, please feel free to reach out to me.

Contact: tanggn@bc.edu