MATH 40024/50024: Computational Statistics Final Project

NEWYORK AIRBNB 2019 DATA ANALYSIS

Paritosh Gandre

December 07, 2023

Introduction [15 points]

- What research question(s) would you like to answer?

1 How is the sampling distribution of mean prices?

2 Linear Model to predict **price** using availability_365 column/variable.

3 Linear Model for each neighbourhood_group using room_num & availability_365 columns.

4 Comparison between Linear and Quadratic Model, and find out which is best?

5 Analyzing model on a **uncertain dataset**.

- Why a data-driven, computational approach may be useful to answer the questions?

The main reason behind a data-driven computational approach is, this approach allows us to efficiently handle large datasets, any discrepency in the dataset, finding out complex patterns hidden into the dataset, with iterative computation. With this approach we can automate a lot of tasks and a scalable method for extracting information from various datasets.

- Describe the dataset that you choose.

So, The dataset I have choosen belongs to New York. It is an Airbnb dataset of New York from 2019. The dataset has **~49,000** observations and 16 variables.

Columns Present in dataset are :

Id : Listing ID of the property

Name : Name of the listed property.

host_id : ID of the property owner.

host_name : Name of the property owner.

neighbourhood_group : Location at which property located.

Neighbourhood : Area in which property located.

Latitude : Latitude coordinate.

Longitude : Longitude coordinate.

room_type : Type of the room (Entire Home/ Appt, Private Room, Shared Room)

Price : Price in Dollars Per night.

Minimum_nights : Amounts of minimum night stay at property

number_of_reviews : No. Of reviews

last_review : last review on which date.

reviews_per_month : Numbers of reviews per months.

calculated_host_listings_count : Count of properties listed on that host.

availability_365 : Number of days when listing is available for booking

Computational Methods [30 points]

- For the choosen dataset, what are the necessary data wrangling steps to make the data ready for subsequent analyses?

Removing the null values, **creating new variables**, and **changing data type of columns**, these are the steps that I feel are necessary data wrangling steps for the choosen dataset.

- What exploratory analyses and modeling techniques can be used to answer the research questions?

Exploratory Analysis:

Descriptive Statistics: For descriptive statistics, summary is including mean, minimum and maximum values are computed.

Visualization: Plots, including histogram, bar graph, and scatter plots to analyze categorical columns and relationships between variables.

Modeling Techniques:

Linear Regression : Linear Regression are extensively used for prediction and also for analyzing relationships between response and predictor variable.

Quadratic Model : Quadratic models are implemented to capture non-linearity between predictor and response variable, which involves quadratic terms.

Bootstrapping : A resampling technique involving sampling a number of times with replacement, often to generate uncertainty, or sampling distribution

- What metrics will be used to evaluate the quality of the data analysis?

Regression Metrics: Metrics like **Mean Absolute Error(MAE)**,**Root Mean Squared Error(RMSE)** can be used to evaluate model performances.

Model Comparison Metrics: For comparing the model **ANOVA**, and **R-Squared** measures can be used.

Bootstrapping Metrics : Analyzing the spread or constructing confidence intervals or Standard errors for $\hat{\beta}_0$ and $\hat{\beta}_1$ values.

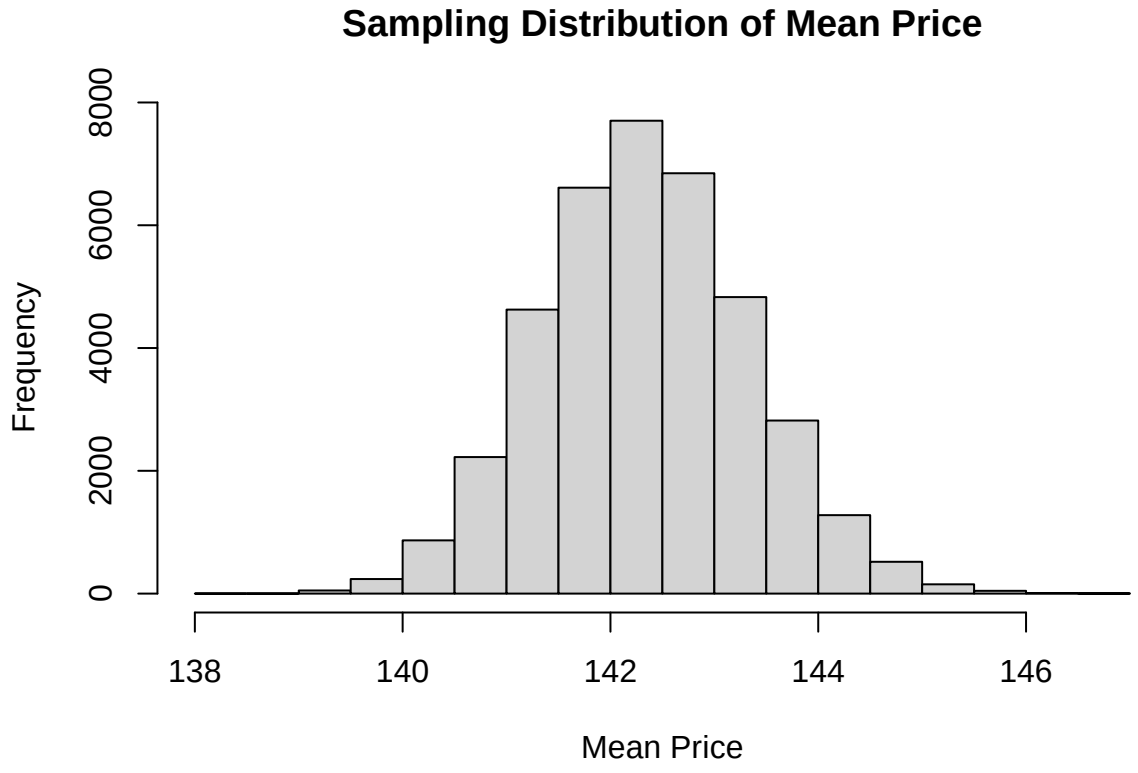
Data Analysis and Results [40 points]

- Perform data analysis, document the analysis procedure, and evaluate the outcomes.
- Present the data analysis results.
- Interpret the results in a way to address the research questions.

importing the data & performing Data Wrangling steps with all the libraries required.

1 How is the sampling distribution of mean prices?

```
# sampling distribution of mean price
set.seed(123)
sampling_means  =  replicate(nrow(ny), mean(sample(ny$price, replace = TRUE)))
hist(sampling_means, main = "Sampling Distribution of Mean Price", xlab = "Mean Price")
```



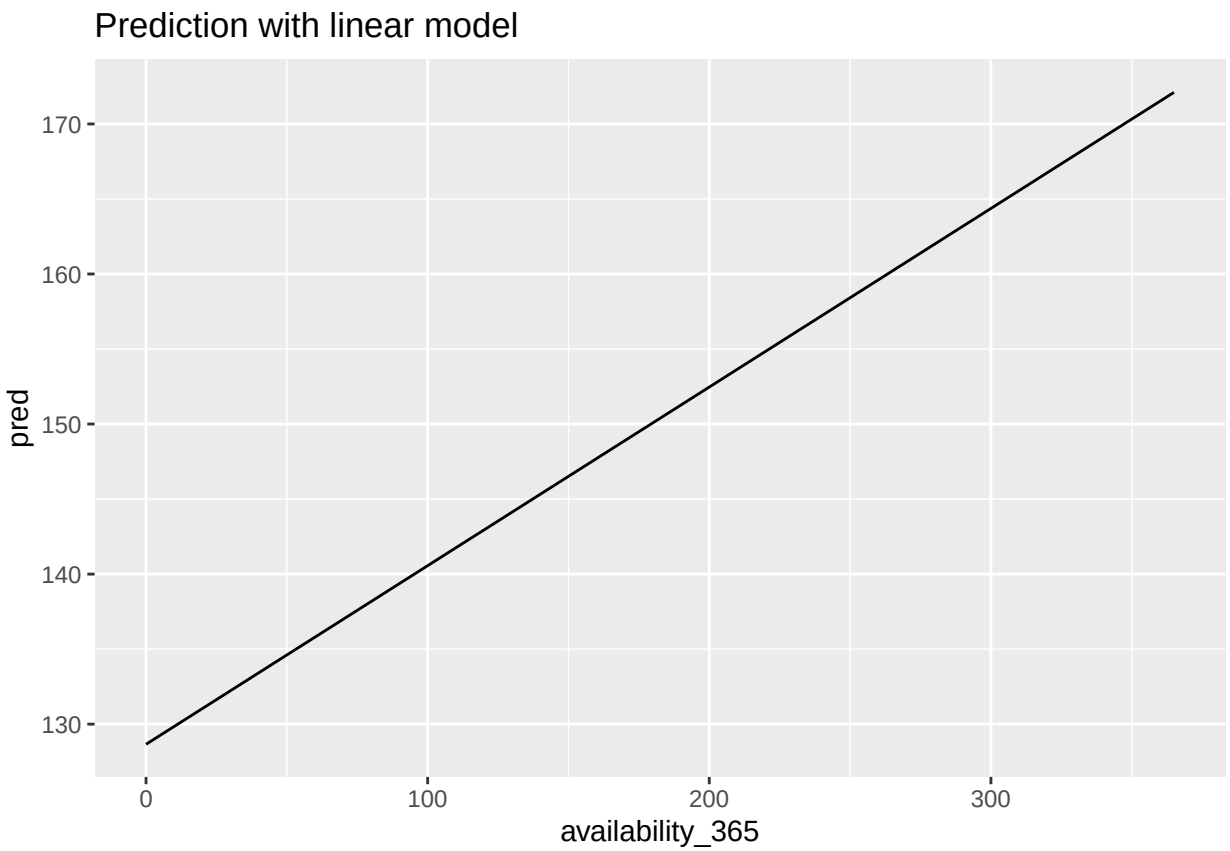
The histogram of the sampling distribution of mean prices provides insights into the variability of mean prices obtained from multiple random samples. The output generated is symmetric and bell-shaped, so we can say that the mean of price is normally distributed.

2 Linear Model to predict **price** using availability_365 column/variable.

```
# Model to predict price using availability_365
new_lm = lm(price ~ availability_365,data = ny)
summary(new_lm)
```

```
##
## Call:
## lm(formula = price ~ availability_365, data = ny)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -168.3  -73.7  -33.7   26.7 9871.3
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.287e+02  1.332e+00   96.56   <2e-16 ***
## availability_365 1.190e-01  7.695e-03   15.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 196.4 on 38819 degrees of freedom
## Multiple R-squared:  0.006127,   Adjusted R-squared:  0.006102
## F-statistic: 239.3 on 1 and 38819 DF,  p-value: < 2.2e-16
```

```r
# Predicted values vs Availability_365
ny |>
  mutate(pred = predict(new_lm)) |>
  ggplot(aes(availability_365, pred)) +
  geom_line() +
  ggtitle("Prediction with linear model")
```



This linear model suggests that there is a statistically significant relationship between the "availability_365" variable and the predicted price. However,the R^2 value is 0.006(0.6%)of the proportion of variance is explained by the model, which makes it relatively low. The summary tells us, we may need to add more variables to the model to improve the predictive performance of the model.

3 Linear Model for each neighbourhood_group using room_num & availability_365 columns.

```
# linear model for each neighbourhood group using room_num and availability_365
all_ng  =  unique(ny$neighbourhood_group)
fits  =  vector("list", length(all_ng))

ny_nested = ny |> nest(data= -c(neighbourhood_group))

neighgroup_fit  =  function(df) {
  lm(price ~ availability_365 + room_num, data = df)
}

ny_nest_lm = ny_nested |> mutate(fit = map(data,neighgroup_fit),
                                 tidied = map(fit,tidy))

# expanding model outputs
lm_result_unnest = ny_nest_lm |> unnest(tidied)
```

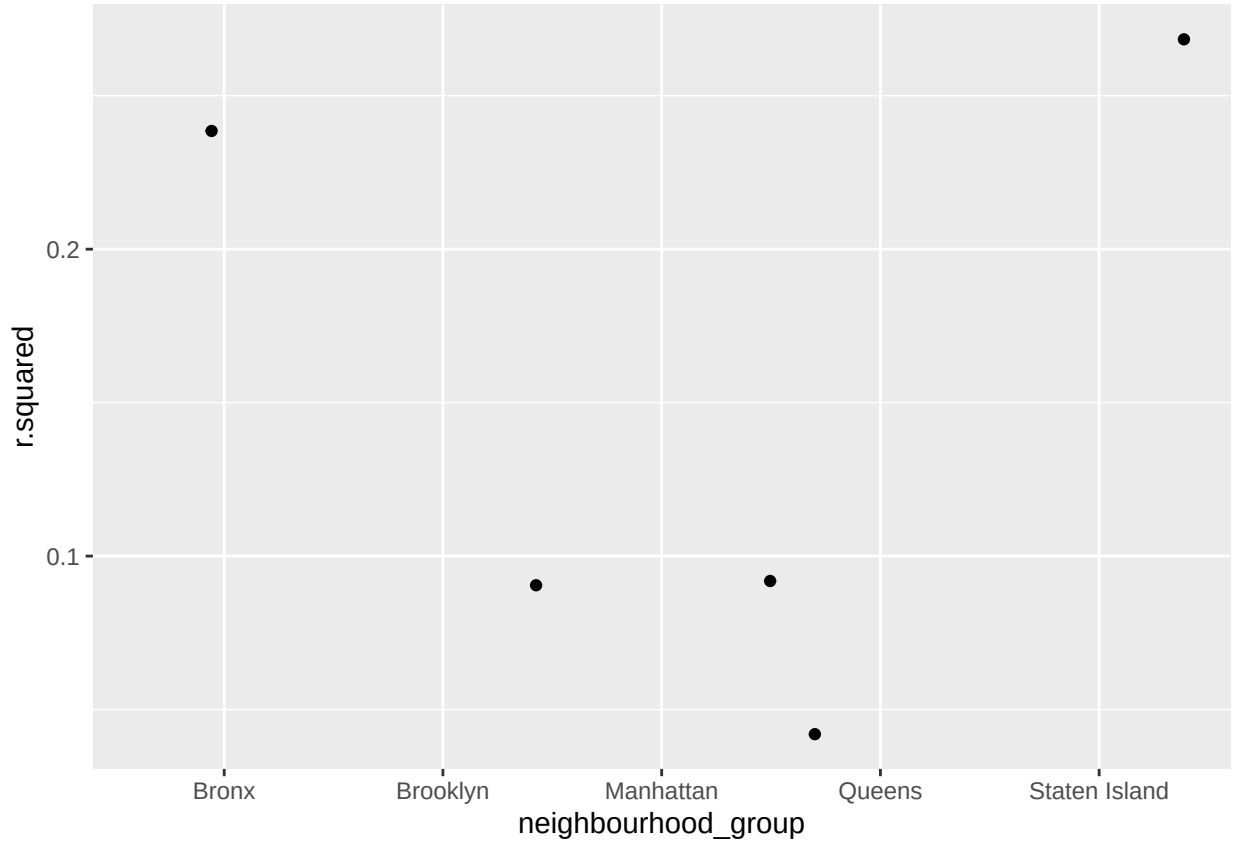
```
lm_result_unnest |>
  filter(neighbourhood_group == "Manhattan")
```

```
## # A tibble: 4 x 8
##   neighbourhood_group data     fit    term          estimate std.error statistic
##   <chr>               <list>   <list> <chr>            <dbl>     <dbl>     <dbl>
## 1 Manhattan           <tibble> <lm>   (Intercept)    204.       2.61       78.2
## 2 Manhattan           <tibble> <lm>   availability~    0.236    0.0131      18.1
## 3 Manhattan           <tibble> <lm>   room_num2     -122.       3.48       -35.1
## 4 Manhattan           <tibble> <lm>   room_num3     -161.      11.7        -13.8
## # i 1 more variable: p.value <dbl>
```

```
# adding residuals
add_resid  =  function(df, model) {
  mutate(df, resid = resid(model))
}

ny_nest_lm = ny_nest_lm |>
  mutate(resids = map2(data,fit,add_resid))

# unnesting a nested data

ny_resids = ny_nest_lm |>
  select(neighbourhood_group,resids) |>
  unnest(resids)

nynnest_glance =  ny_nest_lm |>
  mutate(glance = map(fit, glance)) |>
  select(neighbourhood_group, glance) |>
  unnest(glance)

nynnest_glance |> arrange(r.squared)
```

```
## # A tibble: 5 x 13
##   neighbourhood_group r.squared adj.r.squared sigma statistic  p.value    df
##   <chr>                 <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>
## 1 Queens                0.0422        0.0416 165.       67.1 1.94e-42     3
## 2 Brooklyn              0.0901        0.0900 162.      543.  0            3
## 3 Manhattan             0.0914        0.0913 216.      557.  0            3
## 4 Bronx                 0.239         0.236   55.9      91.2 2.52e-51     3
## 5 Staten Island         0.269         0.261   56.6      37.9 6.58e-21     3
## # i 6 more variables: logLik <dbl>, AIC <dbl>, BIC <dbl>, deviance <dbl>,
## #   df.residual <int>, nobs <int>
```

```
ggplot(nynnest_glance, aes(neighbourhood_group, r.squared)) +
  geom_jitter(width = 0.5)
```



As per the output, the coefficient "availability_365", there is a 0.236 rise in price for every unit of increase in availability in Manhattan.We consider residuals, such as variations between observed and projected values. Within each group, independent linear model fits are possible because to nested data structures. The nynnest_glance tibble presents the overall model assessment, which is arranged according to R-squared values to facilitate comprehension of the predictors' influence and goodness of fit on housing prices among neighborhood groups.

4 Comparison between Linear and Quadratic Model, and find out which is best?

```
# Linear and quadratic model

# Split the data into training and testing sets
set.seed(123)
train_indices  =  sample(1:nrow(ny), 0.8 * nrow(ny))
train_data  =  ny[train_indices, ]
unique(train_data$room_num)
```

```
## [1] 2 1 3
## Levels: 1 2 3
```

```
test_data  =  ny[-train_indices, ]

# Create linear model
linear_model  =  lm(price ~ room_num + availability_365, data = train_data)
```

```r
# Create quadratic model
quadratic_model  =  lm(price ~ room_num + availability_365 + I(room_num)^2
                      +I(availability_365^2), data = train_data)

# Make predictions on the test set
linear_predictions  =  predict(linear_model, newdata = test_data)
quadratic_predictions  =  predict(quadratic_model, newdata = test_data)



# Calculate accuracy-like metric for linear model
linear_accuracy  =  mean(abs(test_data$price - linear_predictions) )

# Calculate accuracy-like metric for quadratic model
quadratic_accuracy  =  mean(abs(test_data$price - quadratic_predictions) )

# Compare model accuracy
cat("Linear Model Accuracy:", linear_accuracy, "\n")
```

```
## Linear Model Accuracy: 66.41096
```

```r
cat("Quadratic Model Accuracy:", quadratic_accuracy, "\n")
```

```
## Quadratic Model Accuracy: 66.42506
```

```r
anova_result = anova(linear_model,quadratic_model)
anova_result
```

```
## Analysis of Variance Table
##
## Model 1: price ~ room_num + availability_365
## Model 2: price ~ room_num + availability_365 + I(room_num)^2 + I(availability_365^2)
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1  31052 1048684035
## 2  31051 1048650097  1     33939 1.0049 0.3161
```

Two models are compared in the ANOVA table: Model 1, which is a simpler linear model with room_num and availability_365 predictors, and Model 2, which is a more complex quadratic model with squared terms for room_num and availability_365. With respect to the F-statistic, the p-value is 0.3161, which is higher than the usual significance level of 0.05.

After looking at the P-value, we do not have enough evidence to reject null hypothesis which is performance for both the models is similar.When both the models have similar performances, we always use the linear regression model. Therefore, we can use linear model for our future computations.

5 Analyzing model on a **uncertain dataset**.

```r
# Creating a new dataset with Uncertainty for prediction

s = 2000
n = nrow(ny)
n
```

```
## [1] 38821
```

```r
boot_intercept = rep(0, s)
boot_slope = rep(0, s)
for (i in 1:s) {
  boot_sample = sample(1:n, replace = TRUE)
  boot_data = ny[boot_sample, ]

  lm_result = lm(price ~ room_num + availability_365, data = boot_data)

  boot_intercept[i] = coef(lm_result)[1]
  boot_slope[i] = coef(lm_result)[2]

}
boot_se_intercept = sd(boot_intercept)
boot_se_slope = sd(boot_slope)
summary(boot_intercept)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   177.2   180.3   181.4   181.4   182.5   186.8
```

```r
summary(boot_slope)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -120.2  -114.3  -113.0  -113.0  -111.7  -106.4
```

```r
summary(linear_model)
```

```
##
## Call:
## lm(formula = price ~ room_num + availability_365, data = train_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -215.1  -54.6  -19.7   16.7 9931.7
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.801e+02  1.697e+00  106.15   <2e-16 ***
## room_num2        -1.118e+02  2.114e+00  -52.89   <2e-16 ***
## room_num3        -1.375e+02  7.246e+00  -18.97   <2e-16 ***
## availability_365  1.354e-01  8.059e-03   16.80   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 183.8 on 31052 degrees of freedom
## Multiple R-squared:  0.0924, Adjusted R-squared:  0.09231
## F-statistic:  1054 on 3 and 31052 DF,  p-value: < 2.2e-16
```

In order to evaluate the model and variability, we apply the model on uncertain dataset. I have created a new dataset using bootstrap method with 2000 samples. The computed **Standard Error** and **Intercept** are 1.96 and 1.53,respectively. The computation showed a mean intercept of 181.4 and a mean slope of -113.0. When the initial linear model was applied to the main dataset, it revealed an intercept of 180.1 and shed light on how various room_num values affected Airbnb prices. With a multiple R-squared of 0.0924, the explanatory power appeared to be moderate. So, this analysis also shows us the importance of considering uncertain dataset.

Conclusion [15 points]

- Does the analysis answer the research questions?
  - Every research question is satisfactorily addressed in the analysis. In addition to constructing a linear model for price prediction based on availability_365, it investigates the sampling distribution of mean prices, compares linear and quadratic models, and assesses model performance on an uncertain dataset. It also builds linear models for various neighborhood groups. Comprehensive responses to each research question are provided by visualizations, summary statistics, and model evaluations, all of which are backed by code outputs and interpretations.
- Discuss the scope and generalizability of the analysis.
  - The scope of the analysis is well-defined within the context of the New York Airbnb dataset from 2019. The findings are relevant to this specific dataset, and the insights gained from neighborhood-specific linear models contribute to understanding housing dynamics in different areas. However, the generalizability of the linear models to other locations or time periods may be limited. The uncertainty analysis enhances the robustness of the linear model, offering valuable insights into parameter variability. To improve generalizability, consideration of more features, modeling techniques could be explored.
- Discuss potential limitations and possibilities for improvement.
  - One of the limitations of the linear model is its low R-squared value, which suggests limited predictive power. Results may be impacted by problems with data quality, and the linearity assumptions should be carefully examined. A more complex understanding of price variations could involve the incorporation of temporal dynamics. In spite of these things, the analysis provides a strong basis and creates opportunities for improvement and growth.