Assignment 1

Lorenzo Ausiello

2023-09-26

PROBLEM 1

Introduction The purpose of this report is to analyze information about real estate sales data in the different boroughs of New York City, taking into account sales from 2012 to 2013. The analysis aims to identify any dependencies or trends in the data and draw conclusions based on the results.

Data Description The datasets include columns such as borough, neighborhood, building class category, tax class at present, block, lot, residential units, commercial units, total units, land square feet, gross square feet, year built, tax class at time of sale, building class at time of sale, sale price, and sale date.

Data Cleaning Once found out outliers, missing values, and values equal to 0, it has been decided beforehand to remove records with no price information. Then, records with outliers (log(square feet)<=4 or log(sale.price)<=6) have been removed, since they affect the accuracy of the analysis. Below table and scatter plot. Column names have been converted to lowercase. Moreover, the dates have been formatted correctly, and it has been giving to the variable useful format. The 5 different datasets, relating to the 5 New York City boroughs, have been merged to simplify analysis, without loss of information.

Data Analysis Below is the statistics table of the sale prices before and after removing missing value and zero values, and then the statistics table after removing outliers. Below is also scatter plot between log(square feet) and log(sale price) before and after.

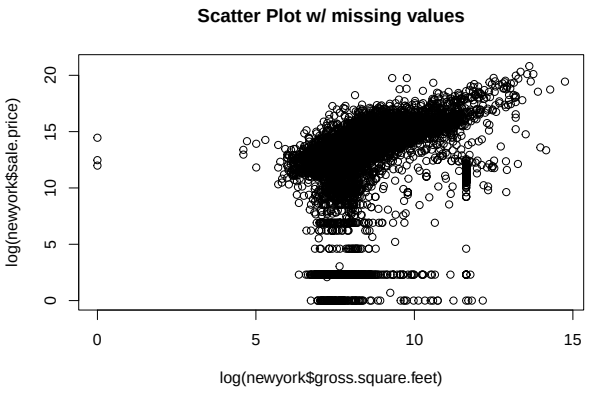


Table 1: Sale Prices Statistics w/ missing values

Min	0.0
Max	1307965050.0
Mean	885097.9

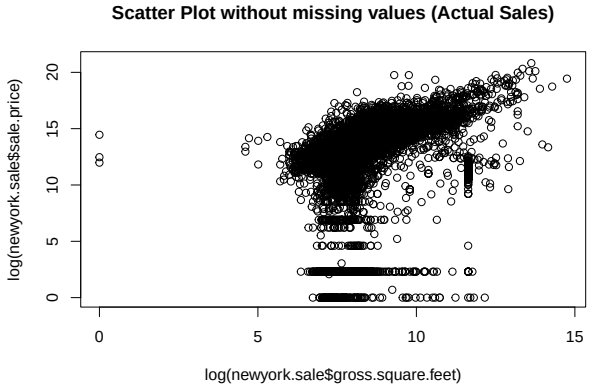


Table 2: Sale Prices Statistics without missing values (Actual Sales)

Min	1
Max	1307965050
Mean	1327176

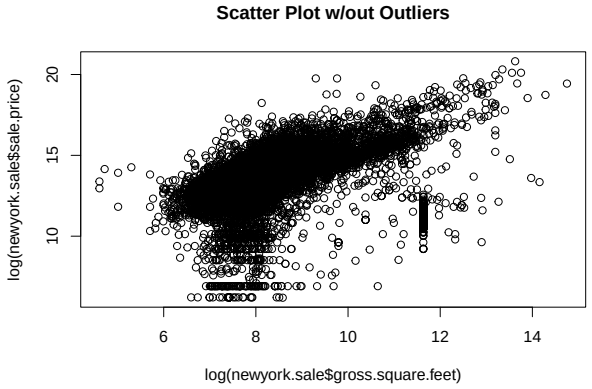
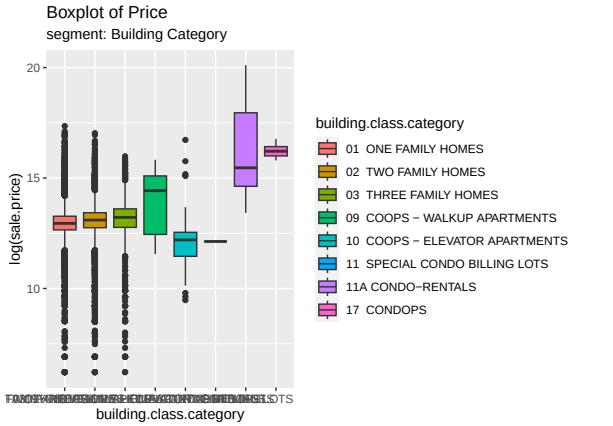


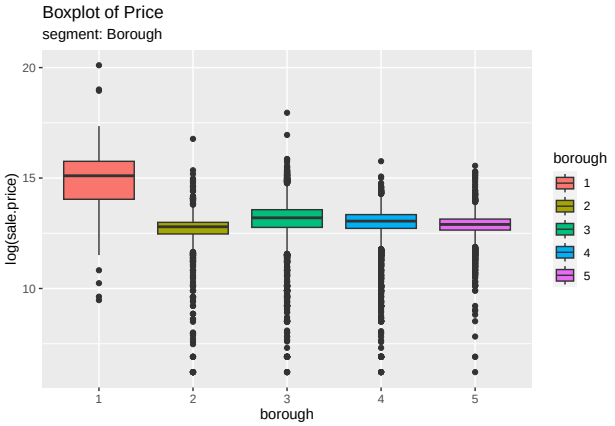
Table 3: Sale Prices Statistics w/out Outliers

Min	490
Max	1100000000
Mean	1661751

The scatter plots seem to show a positive relationship between price and square feet, which means that as the gross square feet increase, the sale price also tends to increase. Positive relationship is more apparent after removing outliers. The x-axis represents the log of the gross square feet, and the y-axis represents the log of the sale price. The log transformation is useful when the data has a wide range of values, as it can help to reduce the effect of outliers and make the relationship between the variables more apparent. Further analysis are needed.

Now let's look only at 1, 2, 3 family homes,coops, and condos. A new variable (sale.datewindow) has been created to categorize records in 5 different quarters based on sale date. Below statistics and graphs.





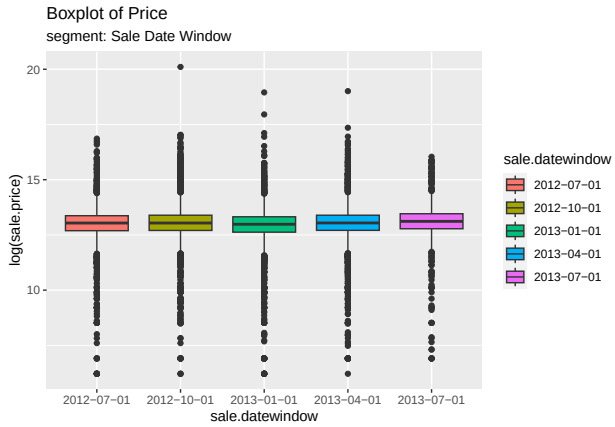


Table 4: Statistics of Sale Price - segment: Building Category

Building Category	Length	Min	Max	Mean
01 ONE FAMILY HOMES	9428	500	34350000	568948.7
02 TWO FAMILY HOMES	7969	500	24912140	601339.6
03 THREE FAMILY HOMES	2163	500	8750000	685632.6
09 COOPS - WALKUP APARTMENTS	17	104000	7500000	2283832.1
10 COOPS - ELEVATOR APARTMENTS	28	13000	18386667	1338854.4
11 SPECIAL CONDO BILLING LOTS	1	185000	185000	185000.0
11A CONDO-RENTALS	13	675000	539000000	75993711.8
17 CONDOPS	5	7250000	19248600	11994720.0

Table 5: Statistics of Sale Price - segment: Borough

Borough	Length	Min	Max	Mean
1	319	13000	539000000	7765968.1
2	1783	500	19248600	385966.3
3	6409	500	62742225	666045.5
4	8143	500	7000000	494397.3
5	2970	500	5700000	438872.8

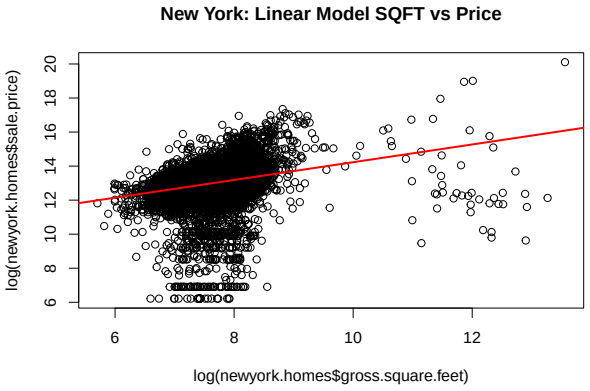
Table 6: Statistics of Sale Price - segment: Sale Date Window

Sale Date Window	Length	Min	Max	Mean
2012-07-01	3309	500	21000000	589289.1
2012-10-01	4880	500	539000000	757316.4
2013-01-01	4543	500	170000000	593945.2
2013-04-01	5241	500	180000000	639565.4
2013-07-01	1651	1000	9245000	646650.4

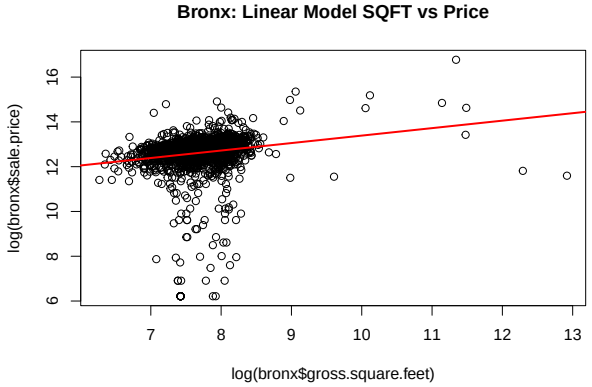
Visualizations were created to explore the relationship between the sale price and other variables, such as borough, neighborhood, and building class category. Looking at boxplots and statistics, CONDOPS is by far the most expensive Building Category, and SPECIAL CONDO BILLING LOTS is the cheapest. As for boroughs, Manhattan is the most expensive real estate area, and it has a mean price that is 20 times higher than the cheapest borough mean price, i.e. Bronx. Furthermore, New York City Real Estate mean prices hit a low during the first quarter of the period considered, and they reached their peak immediately afterwards, during the second quarter.

Finally, linear regressions have been calculated, analyzing the relationship between log(sale prices) (response variable) and log(square feet) (explanatory variables). It has been made for New York City and for each boroughs. Below the results.

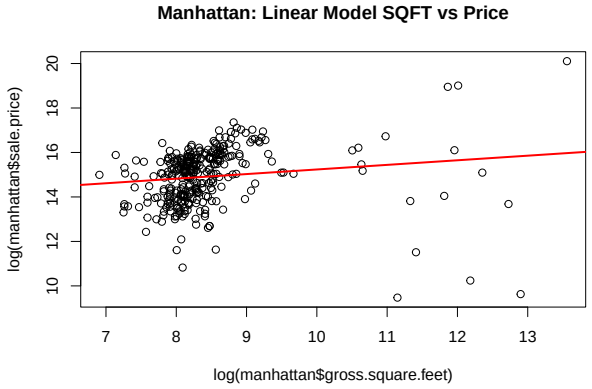
```
##
## Call:
## lm(formula = log(newyork.homes$sale.price) ~ log(newyork.homes$gross.square.feet))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9305 -0.2288  0.0967  0.3981  4.0222
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          9.01468    0.10047   89.72   <2e-16 ***
## log(newyork.homes$gross.square.feet)  0.52133    0.01322   39.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8599 on 19622 degrees of freedom
## Multiple R-squared:  0.07344,    Adjusted R-squared:  0.07339
## F-statistic:  1555 on 1 and 19622 DF,  p-value: < 2.2e-16
```



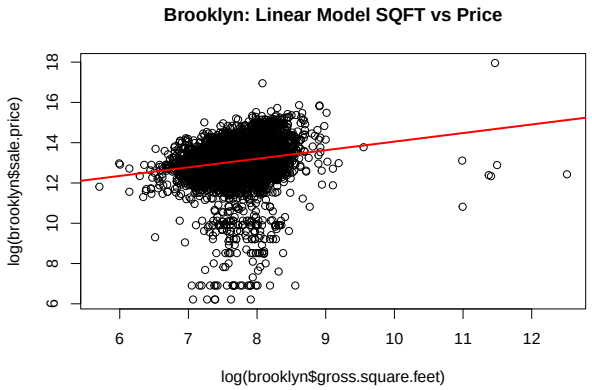
```
##
## Call:
## lm(formula = log(bronx$sale.price) ~ log(bronx$gross.square.feet))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4799 -0.1338  0.2137  0.3840  2.9384
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   10.05233    0.36433  27.591  < 2e-16 ***
## log(bronx$gross.square.feet)  0.33349    0.04734   7.044 2.66e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9203 on 1781 degrees of freedom
## Multiple R-squared:  0.0271, Adjusted R-squared:  0.02656
## F-statistic: 49.62 on 1 and 1781 DF,  p-value: 2.66e-12
```




```
##
## Call:
## lm(formula = log(manhattan$sale.price) ~ log(manhattan$gross.square.feet))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2042 -0.7866  0.2551  0.8495  4.1344
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     13.16732    0.68321  19.273   <2e-16 ***
## log(manhattan$gross.square.feet)  0.20677    0.08049   2.569   0.0107 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.284 on 317 degrees of freedom
## Multiple R-squared:  0.02039,    Adjusted R-squared:  0.0173
## F-statistic:   6.6 on 1 and 317 DF,  p-value: 0.01066
```

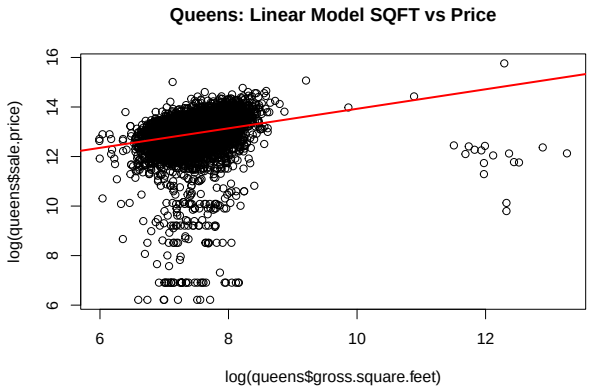



```
##
## Call:
## lm(formula = log(brooklyn$sale.price) ~ log(brooklyn$gross.square.feet))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9473 -0.2900  0.1370  0.4718  3.7166
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     9.79423    0.22354   43.81   <2e-16 ***
## log(brooklyn$gross.square.feet)  0.42584    0.02892   14.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9402 on 6407 degrees of freedom
## Multiple R-squared:  0.03273,    Adjusted R-squared:  0.03258
## F-statistic: 216.8 on 1 and 6407 DF,  p-value: < 2.2e-16
```

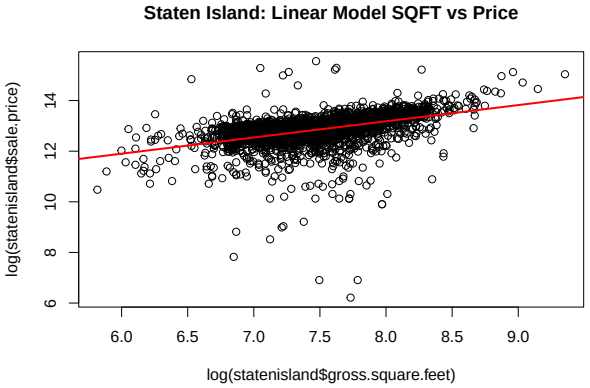



```
##
## Call:
## lm(formula = log(queens$sale.price) ~ log(queens$gross.square.feet))
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -6.8112 -0.1694  0.1293  0.3984  2.2134
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    9.98421    0.14947   66.80   <2e-16 ***
## log(queens$gross.square.feet)  0.39431    0.01998   19.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7719 on 8141 degrees of freedom
## Multiple R-squared:  0.04565,    Adjusted R-squared:  0.04554
## F-statistic: 389.4 on 1 and 8141 DF,  p-value: < 2.2e-16
```




```
##
## Call:
## lm(formula = log(statenisland$sale.price) ~ log(statenisland$gross.square.feet))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7949 -0.1383  0.0806  0.2530  2.7136
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         8.05975    0.17240   46.75   <2e-16 ***
## log(statenisland$gross.square.feet)  0.64017    0.02303   27.80   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5338 on 2968 degrees of freedom
## Multiple R-squared:  0.2066, Adjusted R-squared:  0.2063
## F-statistic: 772.8 on 1 and 2968 DF,  p-value: < 2.2e-16
```



The models reveal a positive relationship between sale prices and square feet in each borough. The p-value associated to the parameters (intercept and slope) are all very close to zero, indicating that the overall model is highly statistically significant. However, the R-squared values are very low. R-Squared, also known as the coefficient of determination, represents the proportion of the variance in the dependent variable (in this case, the logarithm of sale prices) that is explained by the independent variable (in this case, the logarithm of gross square feet). In other words, it measures how well the independent variable can predict or account for the variation in the dependent variable. This suggests that, in each borough, the linear relationship between these two variables, as captured by the model, explains only a small proportion of the total variability in sale prices. This suggests that other factors not included in the model may also influence sale prices.

PROBLEM 2

Introduction The datasets provided nyt1.csv, nyt2.csv and nyt3.csv represent three days of ads shown and clicks recorded on the New York Times homepage. Each row represents a single user. There are 5 columns: age, gender (0=female, 1=male), number impressions, number clicks, and logged-in. We will explore various aspects of the data and perform data analysis.

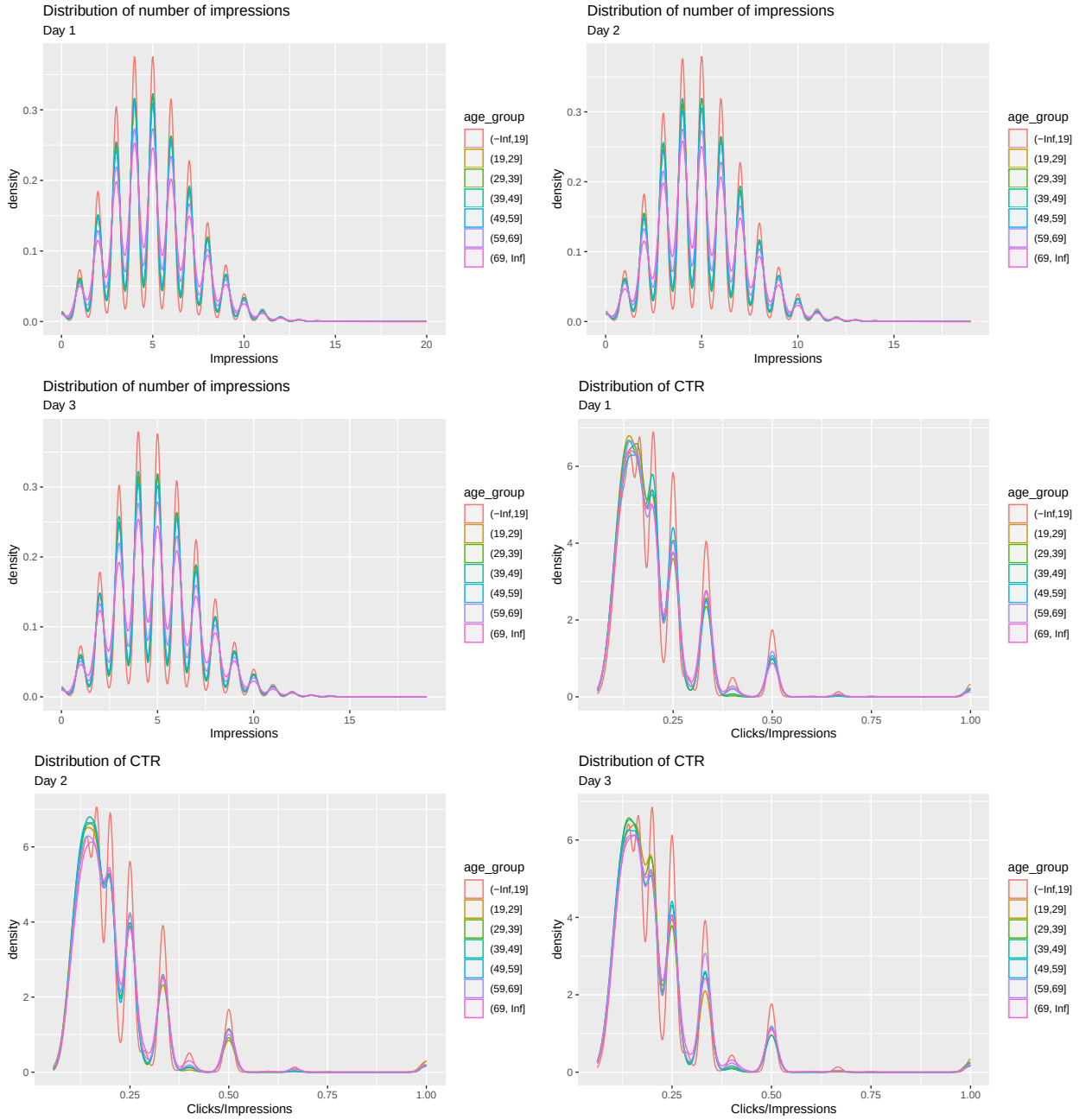
Data Preparation Once read and cleaned the data, giving them useful format, a new variable was created: "age_group". This new variable categorizes users as "<20", "20-29", "30-39", "40-49", "50-59", "60-69", and "70+". The 3 different datasets, relating to the 3 different days, have been merged to simplify starting analysis, without loss of information.

Table 7: Summary All Days

Age	Gender	Impressions	Clicks	Signed_In	age_group	weekday
Min. : 0.00	0:850482	Min. : 0.000	Min. :0.00000	0:403761	(-Inf,19]:479701	1:458441
1st Qu.: 0.00	1:498264	1st Qu.: 3.000	1st Qu.:0.00000	1:944985	(19,29] :169852	2:449935
Median : 31.00	NA	Median : 5.000	Median :0.00000	NA	(29,39] :189026	3:440370
Mean : 29.49	NA	Mean : 5.001	Mean :0.09255	NA	(39,49] :198390	NA
3rd Qu.: 48.00	NA	3rd Qu.: 6.000	3rd Qu.:0.00000	NA	(49,59] :160944	NA
Max. :111.00	NA	Max. :20.000	Max. :6.00000	NA	(59,69] : 95611	NA
NA	NA	NA	NA	NA	(69, Inf]: 55222	NA

The dataset primarily consists of users with a median age of 31, skewed slightly towards younger individuals, and a majority of users are signed in (70.06%). Ad impressions vary from 0 to 20, with a median of 5, while ad clicks are generally low, with a median of 0. Moreover, we are able to notice how the most frequent age group is those under 19s.

Then, the distribution of number of impressions and click-through-rate (CTR = clicks / impressions) for these age categories was plotted, generating the following results:



As shown in the graphs, observations are more frequent for number of Impressions equal to 4 and 5, on all three days and for all age group. The greater the distance from these values, the lower the frequency. As for the distribution of click-through-rate, also in this case are shown similar density frequency function for each day and age group. Particularly, most observations are concentrated between CTR values lower than or equal 0.25, and a few observations corresponding to 0.3 and 0.5.

After that, a new variable was created to segment users based on their click behavior: users with Impressions=0 were included in the "NoImps" segment, users with Impressions>0 but Clicks=0 included in the "Imps" segment, the remaining in "Clicks" category. Subsequently, quantitative and visual comparisons across user segments/demographics were executed.

Table 8: Statistics of Impressions Day 1 - segment: Age Group and Gender

Gender	Age_Group	Min	Max	Mean
0	(-Inf,19]	0	18	5.001456
0	(19,29]	0	16	5.000889
0	(29,39]	0	17	5.018815
0	(39,49]	0	17	5.006110
0	(49,59]	0	15	5.002799
0	(59,69]	0	16	5.033499
0	(69, Inf]	0	15	5.022309
1	(-Inf,19]	0	17	4.988841
1	(19,29]	0	16	4.984637
1	(29,39]	0	20	5.012919
1	(39,49]	0	17	5.022644
1	(49,59]	0	15	5.026173
1	(59,69]	0	16	5.016751
1	(69, Inf]	0	16	4.973289

Table 9: Statistics of Impressions Day 2 - segment: Age Group and Gender

Gender	Age_Group	Min	Max	Mean
0	(-Inf,19]	0	19	5.002963
0	(19,29]	0	16	4.979126
0	(29,39]	0	18	4.973599
0	(39,49]	0	17	4.980391
0	(49,59]	0	16	5.014576
0	(59,69]	0	15	5.003851
0	(69, Inf]	0	16	5.001015
1	(-Inf,19]	0	16	4.987360
1	(19,29]	0	15	5.019542
1	(29,39]	0	16	4.995643
1	(39,49]	0	16	5.004912
1	(49,59]	0	16	5.012958
1	(59,69]	0	16	5.005847
1	(69, Inf]	0	17	4.991627

Table 10: Statistics of Impressions Day 3 - segment: Age Group and Gender

Gender	Age_Group	Min	Max	Mean
0	(-Inf,19]	0	17	4.997024
0	(19,29]	0	17	4.988877
0	(29,39]	0	18	5.010251
0	(39,49]	0	16	4.996967
0	(49,59]	0	16	4.974323
0	(59,69]	0	16	4.991338
0	(69, Inf]	0	15	4.997654
1	(-Inf,19]	0	17	5.003186

Gender	Age_Group	Min	Max	Mean
1	(19,29]	0	18	4.987582
1	(29,39]	0	19	5.008530
1	(39,49]	0	16	4.989990
1	(49,59]	0	16	5.013029
1	(59,69]	0	14	4.995790
1	(69, Inf]	0	16	4.965807

Table 11: Statistics of Impressions Day 1 - segment: Clicks Cat and Signed In

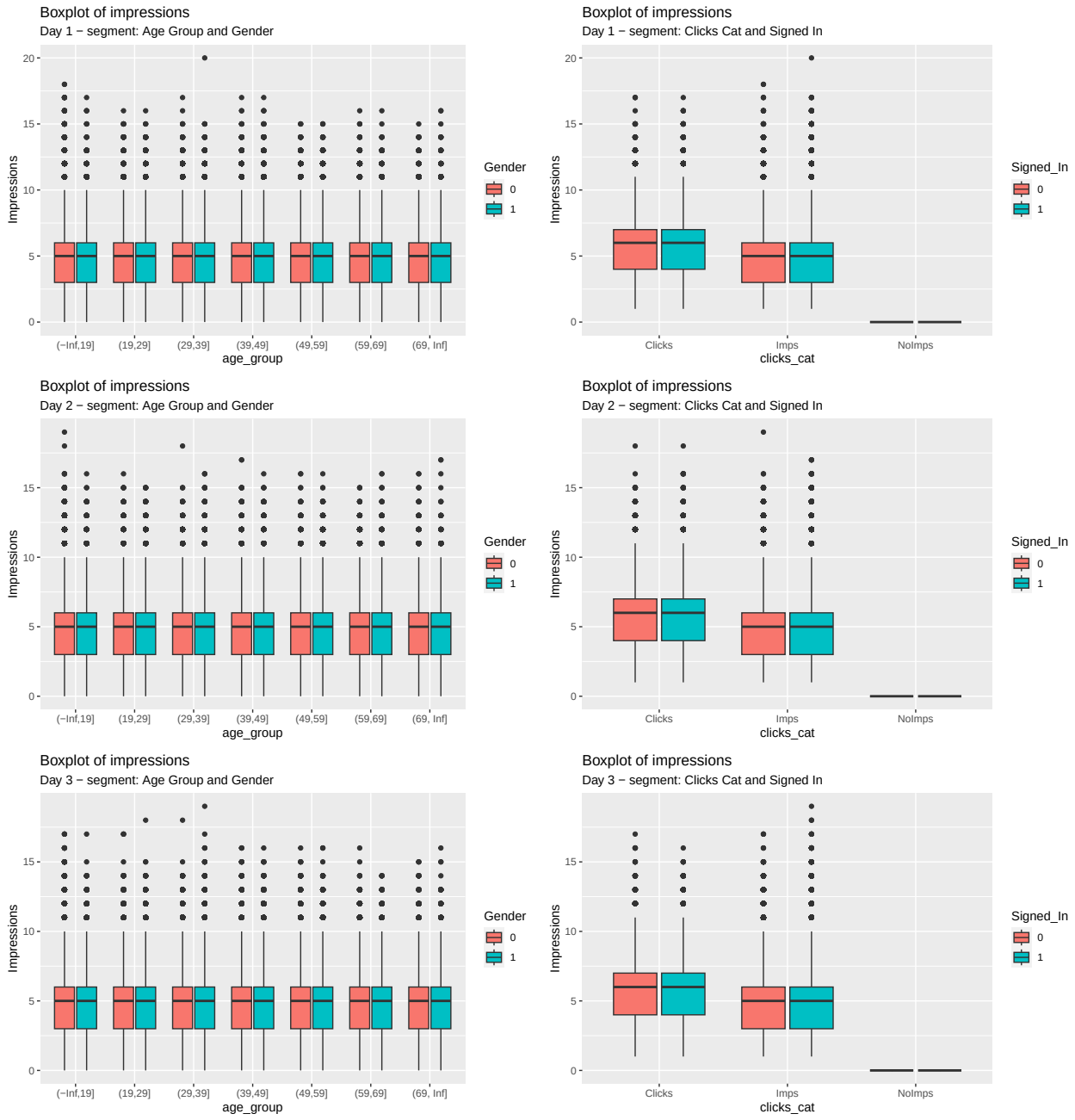
Clicks_Cat	Signed_In	Min	Max	Mean
Clicks	0	1	17	5.922480
Clicks	1	1	17	5.985405
Imps	0	1	18	4.900339
Imps	1	1	20	4.974241
NoImps	0	0	0	0.000000
NoImps	1	0	0	0.000000

Table 12: Statistics of Impressions Day 2 - segment: Clicks Cat and Signed In

Clicks_Cat	Signed_In	Min	Max	Mean
Clicks	0	1	18	5.965146
Clicks	1	1	18	5.952040
Imps	0	1	19	4.897294
Imps	1	1	17	4.963966
NoImps	0	0	0	0.000000
NoImps	1	0	0	0.000000

Table 13: Statistics of Impressions Day 3 - segment: Clicks Cat and Signed In

Clicks_Cat	Signed_In	Min	Max	Mean
Clicks	0	1	17	5.906923
Clicks	1	1	16	5.925709
Imps	0	1	17	4.897480
Imps	1	1	19	4.965281
NoImps	0	0	0	0.000000
NoImps	1	0	0	0.000000



The impressions mean is roughly the same both for the two genders and for the different age groups. It is around a value of 5. Max values among different segments are roughly the same as well: the highest difference (4 impressions) is between 29-39 (18/19 impressions) and over 59 (14/15 impressions). It is interesting to notice that users included in "Clicks" category present a mean value of Impressions higher than users included in "Imps" category. Therefore, users with at least 1 Click are those users who interacted more and so with higher values of Impressions.

Finally, analysis was extended to make further comparisons across days. Below some metrics and distributions over time.

Table 14: Summary Day 1

Age	Gender	Impressions	Clicks	Signed_In	age_group	weekday	clicks_cat
Min. : 0.00	0:290176	Min. : 0.000	Min. :0.00000	0:137106	(-Inf,19]:162867	Min. :1	Clicks: 39838
1st Qu.: 0.00	1:168265	1st Qu.: 3.000	1st Qu.:0.00000	1:321335	(19,29] : 57715	1st Qu.:1	Imps :415537
Median : 31.00	NA	Median : 5.000	Median :0.00000	NA	(29,39] : 64763	Median :1	NoImps: 3066
Mean : 29.48	NA	Mean : 5.007	Mean :0.09259	NA	(39,49] : 67565	Mean :1	NA
3rd Qu.: 48.00	NA	3rd Qu.: 6.000	3rd Qu.:0.00000	NA	(49,59] : 54406	3rd Qu.:1	NA
Max. :108.00	NA	Max. :20.000	Max. :4.00000	NA	(59,69] : 32358	Max. :1	NA
NA	NA	NA	NA	NA	(69, Inf]: 18767	NA	NA

Table 15: Summary Day 2

Age	Gender	Impressions	Clicks	Signed_In	age_group	weekday	clicks_cat
Min. : 0.00	0:281479	Min. : 0	Min. :0.0000	0:134572	(-Inf,19]:160012	Min. :2	Clicks: 39173
1st Qu.: 0.00	1:168456	1st Qu.: 3	1st Qu.:0.0000	1:315363	(19,29] : 56610	1st Qu.:2	Imps :407706
Median : 31.00	NA	Median : 5	Median :0.0000	NA	(29,39] : 62799	Median :2	NoImps: 3056
Mean : 29.51	NA	Mean : 5	Mean :0.0928	NA	(39,49] : 66126	Mean :2	NA
3rd Qu.: 48.00	NA	3rd Qu.: 6	3rd Qu.:0.0000	NA	(49,59] : 53870	3rd Qu.:2	NA
Max. :111.00	NA	Max. :19	Max. :5.0000	NA	(59,69] : 32004	Max. :2	NA
NA	NA	NA	NA	NA	(69, Inf]: 18514	NA	NA

Table 16: Summary Day 3

Age	Gender	Impressions	Clicks	Signed_In	age_group	weekday	clicks_cat
Min. : 0.00	0:278827	Min. : 0.000	Min. :0.00000	0:132083	(-Inf,19]:156822	Min. :3	Clicks: 38132
1st Qu.: 0.00	1:161543	1st Qu.: 3.000	1st Qu.:0.00000	1:308287	(19,29] : 55527	1st Qu.:3	Imps :399239
Median : 31.00	NA	Median : 5.000	Median :0.00000	NA	(29,39] : 61464	Median :3	NoImps: 2999
Mean : 29.47	NA	Mean : 4.996	Mean :0.09226	NA	(39,49] : 64699	Mean :3	NA
3rd Qu.: 48.00	NA	3rd Qu.: 6.000	3rd Qu.:0.00000	NA	(49,59] : 52668	3rd Qu.:3	NA

Age	Gender	Impressions	Clicks	Signed_In	age_group	weekday	clicks_cat
Max. :109.00	NA	Max. :19.000	Max. :6.00000	NA	(59,69] : 31249	Max. :3	NA
NA	NA	NA	NA	NA	(69, Inf]: 17941	NA	NA

Table 17: Statistics of Impressions - segment: Weekday and Signed In

Weekday	Signed In	Min	Max	Mean
1	0	0	18	4.999657
1	1	0	20	5.010584
2	0	0	19	5.002772
2	1	0	18	4.998218
3	0	0	17	4.995155
3	1	0	19	4.996850

Table 18: Statistics of Impressions - segment: Weekday and Clicks Cat

Weekday	Clicks Cat	Min	Max	Mean
1	Clicks	1	17	5.957327
1	Imps	1	20	4.953183
1	NoImps	0	0	0.000000
2	Clicks	1	18	5.957905
2	Imps	1	19	4.944978
2	NoImps	0	0	0.000000
3	Clicks	1	17	5.917261
3	Imps	1	19	4.945915
3	NoImps	0	0	0.000000

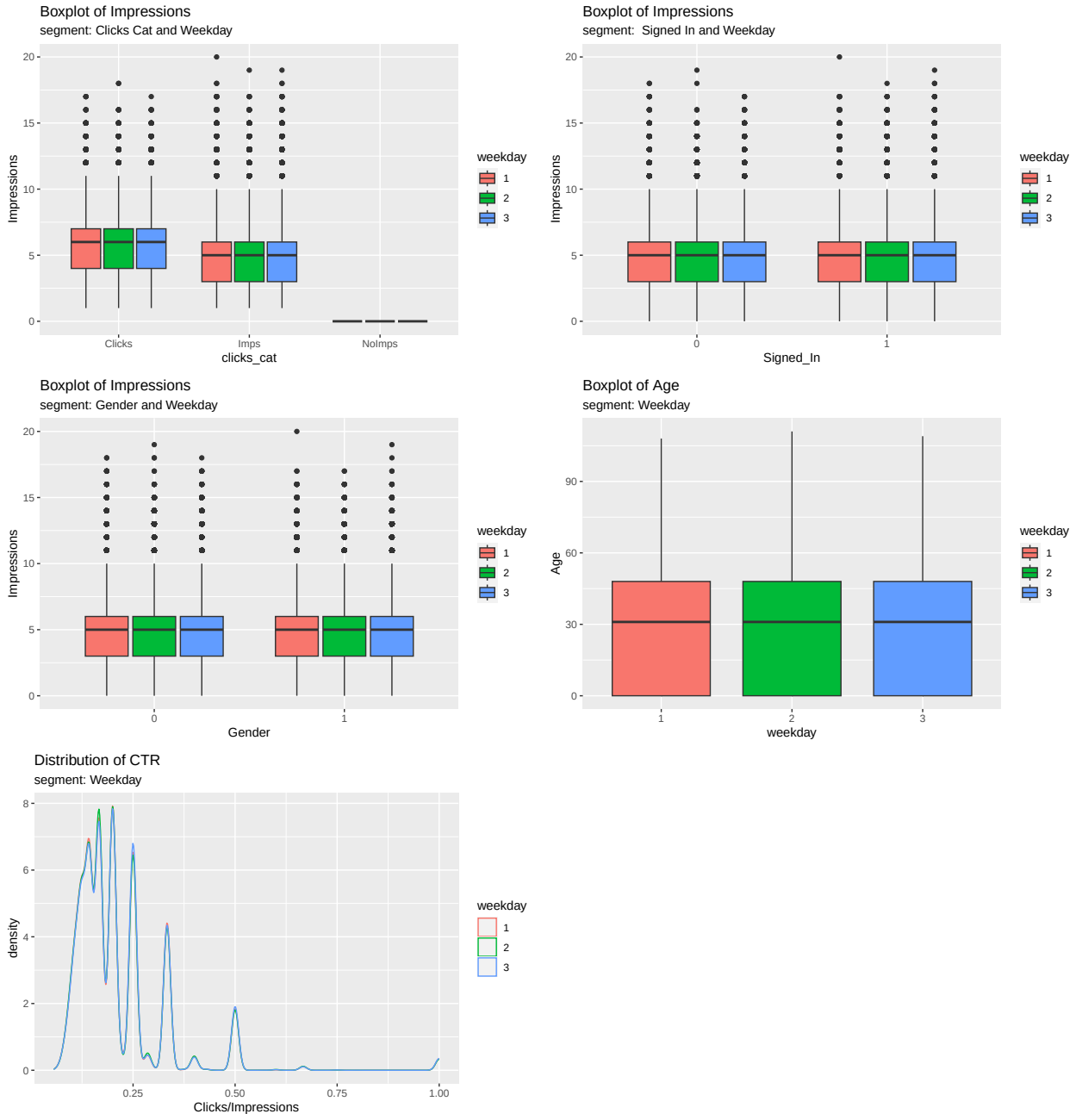
Table 19: Statistics of Impressions - segment: Weekday and Gender

Weekday	Gender	Min	Max	Mean
1	0	0	18	5.006555
1	1	0	20	5.008629
2	0	0	19	4.996348
2	1	0	17	5.004981
3	0	0	18	4.995316
3	1	0	19	4.998112

Table 20: Statistics of Age based on weekday

Weekday	Min	Max	Mean
1	0	108	29.48255
2	0	111	29.50851

Weekday	Min	Max	Mean
3	0	109	29.46802



During the all three days, the impressions means remain constant for the two genders, for the different age groups, for the different clicks categories and also for the different signed in categories. Moreover, also the distribution of CTR remains unchanged during the days.