HYPOTHESIS TESTING

INTRODUCTION

As Data Scientists working for the Autolib electric car-sharing service company we have been tasked with investigating a claim about the average number of bluecars rented out in the city of Paris during the weekdays. Joanneum Research Company places the average number of Bluecars rented out in the city of Paris at 8000 per day. We will source my data from our database currently with daily input of rented cars from January to June to test this claim. We will randomly sample days from our dataset to use in this experiment. we will be using a Z-test of mean since my sample will be more than 30.

HYPOTHESIS

We are claiming that the average number of bluecars rented out on weekdays is less than the average number stated by Joanneum Research company.

Ho: mu >/= 8000.
H1: mu < 8000

I.e

Ho: The average number of bluecars rented out by Autolib Scheme is greater than or equal to 8000

H1: The average number of bluecars rented out by Autolib Scheme is less than to 8000

DATA ANALYSIS

We collected our data from our database of recorded inputs of rented out cars from January to June 2018.

We used the formula below to get my sampling size of 86 days:

$n = N*X / (X + N - 1)$

Where,

$X = Z_{\alpha/2}^{2} *p*(1-p) / MOE^{2}$,

and $Z_{\alpha/2}$ is the critical value of the Normal distribution at α/2 (e.g. for a confidence level of 95%, α is 0.05 and the critical value is 1.96), MOE is the margin of error, p is the sample

proportion, and N is the population size. Note that a Finite Population Correction has been applied to the sample size formula.

Below is my sample:

INDEX	DATE	BLUECARS_RETURNED_SUM	BLUECARS_TAKEN_SUM	SLOTS_FREED_SUM	SLOTS_TAKEN_SUM
52	3-23-2018	8802	8781	2460	2453
37	02-06-18	7631	7732	1985	1965
87	5-22-2018	8399	8489	2278	2253
33	2-22-2018	7723	7676	2100	2105
19	01-04-18	6564	6604	1615	1624
30	02-02-18	8892	8840	2412	2408
73	4-23-2018	6374	6314	1583	1611
0	01-01-18	8144	8098	1906	1897
59	03-06-18	7954	8134	2166	2143
105	6-18-2018	7257	7326	1894	1910
41	03-01-18	7720	7711	2044	2037
108	06-05-18	8456	8651	2266	2230
4	1-15-2018	7451	7443	1820	1790
34	2-23-2018	8005	8079	2109	2109

60	03-07-18	8109	8163	2267	2277
49	3-20-2018	7567	7594	2024	2032
92	5-29-2018	8068	8239	2175	2179
102	6-13-2018	8590	8514	2367	2379
62	03-09-18	8755	8768	2348	2350
43	3-13-2018	7727	7836	1957	1941
55	3-28-2018	8134	8292	2255	2249
39	02-08-18	7120	7122	2078	2064
5	1-16-2018	7718	7851	2043	2050
26	2-14-2018	8222	8375	2381	2352
51	3-22-2018	7990	8054	2133	2108
16	01-03-18	7386	7475	1789	1768
35	2-28-2018	7202	7258	1843	1881
64	04-11-18	8005	8050	2400	2403
11	1-23-2018	7446	7550	1909	1900
29	2-19-2018	6681	6590	1714	1745
8	1-19-2018	9030	9089	2353	2369
6	1-17-2018	8095	8171	2037	2043

57	3-30-2018	8561	8597	2193	2205
13	1-25-2018	7807	7895	2118	2110
84	05-01-18	8869	8707	2159	2190
15	1-29-2018	7333	7313	1975	1999
98	05-08-18	8579	8615	2315	2322
95	5-31-2018	8444	8462	2345	2338
14	1-26-2018	8539	8588	2297	2294
94	5-30-2018	8538	8587	2254	2227
80	04-04-18	7946	8139	2144	2100
74	4-24-2018	6971	7055	1787	1775
42	03-12-18	7232	7241	1949	1948
67	4-16-2018	7149	7071	1870	1854
24	02-12-18	7134	7125	2089	2083
90	5-25-2018	9216	9268	2597	2614
82	04-06-18	8219	8247	2051	2074
3	01-12-18	8609	8668	2104	2100
9	01-02-18	6542	6580	1582	1601
75	4-25-2018	7434	7537	1916	1908

31	2-20-2018	7191	7317	1984	1975
88	5-23-2018	6486	6749	1723	1694
40	02-09-18	7566	7624	2057	2052
68	4-17-2018	7063	7197	1969	1960
2	01-11-18	7807	7738	1917	1928
69	4-18-2018	7289	7421	1971	1974
50	3-21-2018	8005	8062	2132	2135
78	04-03-18	7642	7704	1909	1916
54	3-27-2018	7745	7796	2148	2149
100	06-11-18	8359	8498	2221	2205
56	3-29-2018	8301	8264	2230	2227
25	2-13-2018	7280	7388	2178	2194
48	03-02-18	8400	8357	2248	2265
77	4-27-2018	8015	7980	2003	1984
1	01-10-18	7565	7695	1883	1855
28	2-16-2018	8755	8812	2496	2490
86	5-21-2018	8780	8576	2295	2340
63	04-10-18	7836	8011	2216	2190

23	02-01-18	8169	8289	2242	2239
45	3-15-2018	7794	7818	2143	2100
32	2-21-2018	7458	7549	1898	1882
83	04-09-18	7704	7736	2112	2124
79	4-30-2018	8206	8253	2160	2173
101	06-12-18	8576	8698	2394	2378
17	1-30-2018	7536	7649	2060	2052
109	06-06-18	8423	8506	2229	2253
91	5-28-2018	8123	8265	2062	2036
70	4-19-2018	6569	6666	1677	1640
18	1-31-2018	8136	8160	2170	2169
10	1-22-2018	7737	7725	1959	1953
104	6-15-2018	9141	9226	2466	2466
66	4-13-2018	8781	8787	2475	2484
81	04-05-18	8229	8277	2077	2086
76	4-26-2018	7455	7510	1932	1928
58	03-05-18	7266	7279	1889	1883
27	2-15-2018	8413	8407	2311	2317

There were outliers in our data, we identified them using a boxplot and a histogram during our univariate analysis. We dealt with them by removing them. Our sampled data did not have any missing data or duplicated data. We dropped ['UTILIB_TAKEN_SUM', 'UTILIB_RETURNED_SUM', 'UTILIB_14_TAKEN_SUM', 'UTILIB_14_RETURNED_SUM', 'DAYOFWEEK', 'N_DAILY_DATA_POINTS', 'POSTAL CODE'] columns from our dataset since they were not necessary for our experiment.

Below is a description of the fields we used when collecting our data:

Column name	explanation
Postal code	postal code of the area (in Paris)
date	date of the row aggregation
n_daily_data_points	number of daily data poinst that were available for aggregation, that day
dayOfWeek	identifier of weekday (0: Monday -> 6: Sunday)
day_type	weekday or weekend
BlueCars_taken_sum	Number of bluecars taken that date in that area
BlueCars_returned_sum	Number of bluecars returned that date in that area
Utilib_taken_sum	Number of Utilib taken that date in that area
Utilib_returned_sum	Number of Utilib returned that date in that area
Utilib_14_taken_sum	Number of Utilib 1.4 taken that date in that area
Utilib_14_returned_sum	Number of Utilib 1.4 returned that date in that area
Slots_freed_sum	Number of recharging slots released that date in that area
Slots_taken_sum	Number of rechargign slots taken that date in that area

Below is a description of our sampled data:

count 86.000000
mean 7945.174419
std 623.526521
min 6314.000000
25% 7561.000000
50% 7957.500000
75% 8371.250000
max 9268.000000

Our sample ranges from 6314 to 9268 rented blue cars, with a mean of 7945 and a standard deviation of 623.

We filtered the city of paris using postal codes.since according to wikipedia Paris postal codes start from 75000- 76000, we extracted postal codes in this range.

We also filtered out weekdays from our dataset.

We finally grouped our DATE data using a pivot table in order to obtain a sum of Blue cars taken per day.

Our main challenge is that we only data from some months of one year, which may be affected by seasonality or change of taste and preference in consumers, this may affect the quality of our experiment.

HYPOTHESIS TESTING

Z-test for mean

Step 1 : State the hypothesis and identify the claim.

We are claiming that the average number of bluecars rented out on weekdays is less than the average number stated by Joanneum Research company.

Ho: mu >/= 8000.
H1: mu < 8000

I.e

Null hypothesis
Ho: The average number of bluecars rented out by Autolib Scheme is greater than or equal to 8000

Alternative hypothesis
H1: The average number of bluecars rented out by Autolib Scheme is less than to 8000

Step 2: Find the critical value

At alpha = 0.05 and critical value is -1.645

Step 3: Compute the sample test value.

Sample mean = 7945.174419, **sample standard deviation** = 623.526521

Z = (7945.174419 - 8000)/(623.526521/sqrt(86)) = -0.8154128311791786

Step 4 :lookup the p_value

Probability = 0.20741802105517554

Step 5: Make the decision to reject or not reject the null hypothesis.

We will **reject the alternative hypothesis** in favour of the null hypothesis

0.20741802105517554 **>** 0.05 our alpha value

Step 6: Summarize the results.

There is **not enough evidence to support the claim** that the number of blue cars rented out in paris is less than to 8000

SUMMARY

Our claim lacked enough evidence to prove that the number of Blue cars rented on weekdays was less than the average number as stated by Joanneum Research company. The average rented out blue cars as stated by Joanneum Research company was 8000 cars. I thought that was exaggerated since travel during weekdays is limited but to the contrary the claim was wrong.

Our **test statistics** was **-0.8154128311791786**, which was **higher than our critical value** of -1.645 and lies within a **confidence interval of between 1.645 and 1.645**, hence this made us reject the alternative hypothesis hence discrediting our claim.

Our **p_value** is **0.20741802105517554** which is higher than 0.05 confidence level, infomed the decision to reject our claim.

Our probability of number of rented out cars being less than 8000 is `0.7925819789448245` Which is less than our 0.95 confidence level, hence we are right to discredit this claim.