Programming for Data Science with Python NanoDegree

Project 1

Dirk Mueller
2020-09-19

Length of rental per genre (query_1.sql)

Research question: is there a statistically significant difference in the length of rental between the genre that has the shortest and the one that has the longest mean rental time ?

genre character v	mean_len_rental.. numeric	len_rental_variance.. numeric	n bigint
Sports	5.20	6.84	1179
Travel	4.81	6.64	837

(query_1)

Student's t-test

$$t' = \frac{\mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{5.20 - 4.81}{\text{SQRT}(6.84/1179 + 6.64/837)}$$

$$= 3.33$$

where μ_1 and μ_2 are the mean values, s_1^2 and s_2^2 the sample variances of Sports and Travel, respectively.

ANSWER: yes, there is. Using the student's t-test we infer that the result excedes the critical value – extracted from the t table (two-tailed at 0.05) – which is 1.96.

The null hypothesis, that there is no difference between the samples can be rejected. Hence, there is a statistically significant difference in the length of rentals.

Possible reasons:
1. sport fanatics tend to watch games repeatedly
2. watching a travel documentary might not invoke an interest to rerun the DVD.

*) I checked upfront that the rental and return dates were trustworthy and never null

Customer behavior in late returns (query_2*.sql)

Research question: is there a difference in the amount of rentals returned after the deadline between the genres used in query 1, which were Sports and Travel, in comparison to all genres?

Table 1:

status text	film_count bigint
before deadline	7738
late	6403
on time	1720

(query_2_1)

Table 2:

genre character varying (25)	status text	film_count bigint
Sports	before deadline	516
Sports	late	529
Sports	on time	119

(query_2_sports)

Table 3:

genre character varying (25)	status text	film_count bigint
Travel	before deadline	473
Travel	late	259
Travel	on time	95

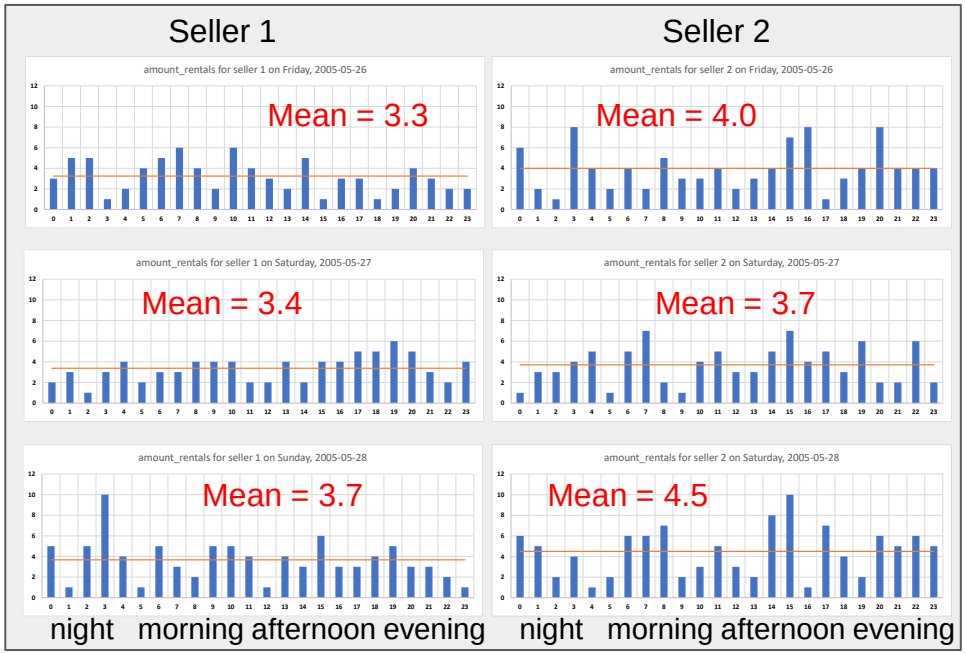
(query_2_travel)

ANSWER: yes, there is.

- 40% of all rentals are returned late (Table 1)
- 45% of Sports rentals are returned late (Table 2)
- 31% of Travel rentals are returned late (Table 3)
- the amount of Sports DVDs returned late is subtantially higher than overall
- the rental company could increase the penalty for returning Sports rentals late

Sellers' performance on a typical weekend (query_3*.sql)

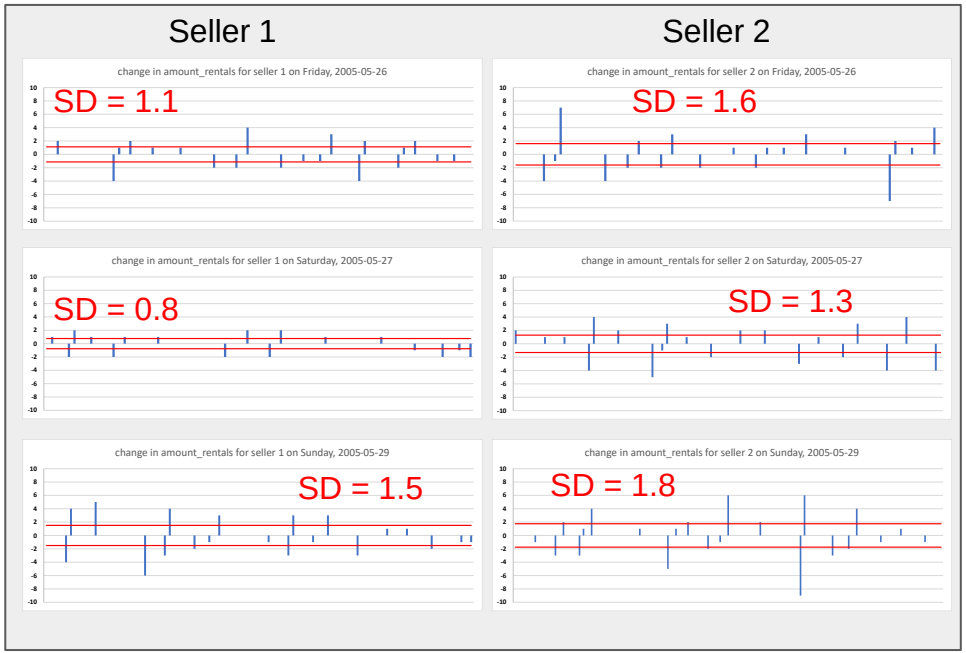
Research question: analyze the rental activity on an hourly basis and which of the two sellers generate more rentals over a typical weekend from Friday, 2005-05-27 to Sunday, 2005-05-29



- seller 2 performs better than seller 1
- Sunday has the highest amount of rentals, Saturday the lowest
- seller 2 is doing very well on Friday and Sunday evenings
- seller 1 is doing poorly on Friday and Sunday evenings
- seller 2 seems to have difficulties around 2-3 a.m.
- overall it seems both sellers complement each other (perhaps they swap the duties regularly)

Variation in Sellers' performance (query_4*.sql)

Research question: what is the fluctuation of rental activity on an hourly basis for the two sellers in terms of rentals over a typical weekend from Friday, 2005-05-27 to Sunday, 2005-05-29?



- rental activities by seller 2 fluctuate much more than for seller 1
- fluctuation in rentals is high on Fridays and Saturdays → some customers might skip the rental because of long lines
- the attention might be impaired by the sellers when things are not moving