Didi

Didi Business Intelligence Challenge

Job position: Business Inteligence

Jesús Alberto Urrutia Camacho

Ciudad de México

26 de junio de 2022

Índice

1. Tasks before making query.

I had to consume cvs files in a database, that I created previously. The database management systems that I used, for this task, is PostgreSQL. I did 3 activities:

- A new database was created, whose name is: test
- Tables was created
- Primary and Foreign Keys (and other constraints) was designed, in order to make possible query. An example is leave in order to show you the SQL query that was used:

```
ALTER TABLE public.restaurant
ADD CONSTRAINT fk_restaurant
FOREIGN KEY (visit_date)
REFERENCES date (calendar_date);
```

2. Write the SQL queries necessary to generate a list of the five restaurants that have the highest average number of visitors on holidays. The result table should also contain that average per restaurant

In order to make querys, I used ANSI standard.

```
SELECT sub.id AS restaurant,
AVG (sub.reserve_visitors) AS promedio
FROM
(
    SELECT da.calendar_date, da.holiday_flg, res.id, res.reserve_visitors
    FROM public.date AS da
    INNER JOIN public.restaurant AS res ON da.calendar_date = res.visit_date
    WHERE da.holiday_flg = 1
    ) AS sub
GROUP BY sub.ID
ORDER BY promedio DESC
LIMIT 5
;
```

3. Use SQL to discover which day of the week there are usually more visitors on average in restaurants.

```
SELECT
sub.dia
FROM
(
SELECT
    da.day_of_week AS dia,
    res.id AS rest,
```

```
    AVG(res.reserve_visitors) AS promedio
    FROM public.date AS da
    INNER JOIN public.restaurant AS res ON da.calendar_date = res.visit_date
    GROUP BY res.id, da.day_of_week
    ORDER BY promedio DESC
    LIMIT 1
    ) AS sub
    ;
```

4. How was the percentage of growth of the amount of visitors week over week for the last four weeks of the data? Use SQL too

```
WITH mensual AS
(
SELECT
    DATE_TRUNC('week', res.visit_datetime) AS SEM,
    SUM(res.reserve_visitors) AS VISITANTE
FROM
    public.date AS da
    INNER JOIN
        public.restaurant AS res    ON da.calendar_date = res.visit_date
    WHERE DATE_TRUNC('week', res.visit_datetime) > '2017-04-24 00:00:00'
    GROUP BY 1
    ORDER BY 1
    )
    SELECT * ,
        LAG(VISITANTE) OVER(ORDER BY SEM) AS INIC,
         (VISITANTE - LAG(VISITANTE) OVER(ORDER BY SEM)) as RESTA,
         (VISITANTE - LAG(VISITANTE) OVER(ORDER BY SEM))/ (LAG(VISITANTE) OVER(ORDER BY SEM))
    FROM mensual
    ;
```

5. Forecast for the next six months, after the last date of the data, the sum of visitors of all the restaurants and validate the accuracy of your forecast. You can solve this question using the tool that you prefer

Due to the fact that one want to model count events (the sum of visitors of all the restaurants), I decided to use a multiple linear regression. However, I give a plus, and I designed a Poisson regression in order to model a counting phenomena.

```
getwd()
setwd("C:/Users/COMIMSA/Documents/aprendizaje/Didi")
bd <- read.csv("new2data-1656198820859.csv", header = T, sep = ",", encoding = "UTF-8")
unique(bd$rest) #En esta base hay 38 restaurantes. No todos los restaurantes se encuentran en los inter
Mode <- function(x) {
  ux <- unique(x)
```

```
  ux[which.max(tabulate(match(x, ux)))]
}
Mode(bd$suma)
mean(bd$suma)
```

A briefly descriptive statistics is done. It is seen that the model's mode is 1 persons per restaurant montly, but the mean is much more higher: 8.15. Thus, the left tail is long.

5.1. Linear Regression

In addition, restaurant, day of the week, holiday, genre, and location, they all are categorical data, but the datatime could be a numerical one. In this scenario, a lineal regression could only use the datatime variable, or turn them all variables into a factors (factor means that a number is just a tag). Thus, the next chunck of code turn the *visit_datetime* variable into a numeric one.

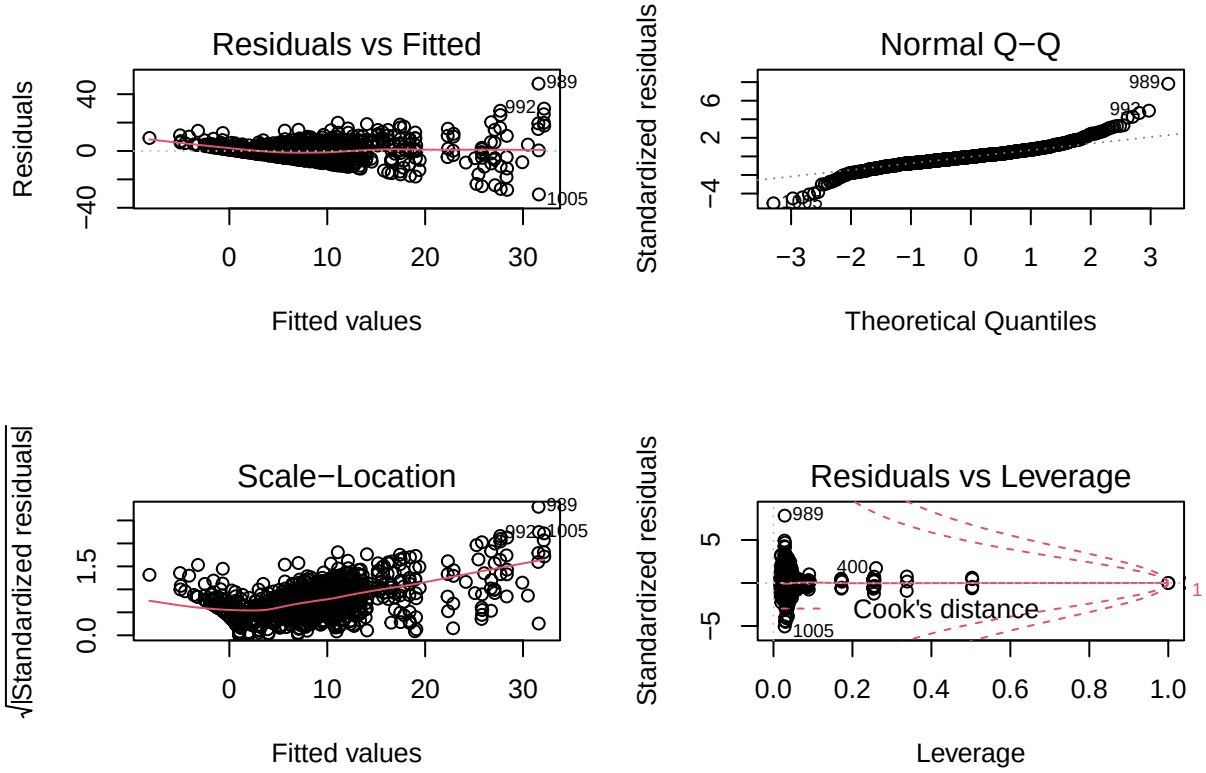
```
bd$mes <- stringr::str_replace_all(bd$mes, "[:punct:]", "")
bd$mes <- gsub(" .*$", "", bd$mes)
bd$mes <- as.integer(bd$mes)
sort(bd$mes, decreasing = T)
str(bd)
```

```
m3 <- lm(suma ~ ., data = bd)
summary(m3)
```

After applying a multiple regression with all the independent variables, the model has a 50 % of statistical significance. Even thought, the intercept is significant, is near zero. Also, some partial regression coefficients are significant, such as those for the day of the week and month. On the other hand, the residuals range is from -30 to 47, which can be considered a wide margin. This could be due to specification errors of the variables that describe the model.

Regarding the theoritical assumptions, a graphical description is shown below:

```
par(mfrow=c(2,2))
plot(m3) #homoscedasticidad pedo
```




```
bptest(m3) #Homoscedast
```

```
##
##  studentized Breusch-Pagan test
##
## data:  m3
## BP = 321.73, df = 45, p-value < 2.2e-16
```

```
dwtest(m3)
```

```
##
##  Durbin-Watson test
##
## data:  m3
## DW = 1.6743, p-value = 0.7377
## alternative hypothesis: true autocorrelation is greater than 0
```

It seems that they are not fulfilled, with the exception of the influential outliers. Although, It looks like if there is dependence between residuals, homosceasticity, and the tails of the normal are very 'heavy'.

Analytical tests do not reject the null hypothesis of heteroscedasticity. Also, it is stated that there is a correlation between residuals. In this scenario, all of these problems may be given by specification errors, or linear dependency.

Thus, It is strongly to propose a model that better fits and enforces the assumptions. Three tools are used to achieve the best model: Backward, forward and regsubsets. **However, the best way to model and predict is using expert knowledge, and collaboration with people who are masters in the field.**

```
#ols_step_forward_p(m3 , details = T)
```

The forward method selects as relevant variables the Restaurate, then the day of the week, finally if it is a holiday. The variables of location or type of food are not relevant for this criterion. However, with the 4 variables selected, the model has low significance (adjusted R of 0.52)

```
#ols_step_backward_p(m3, details = T)
```

The backward model did not remove any variables from all the initials. But the statistical significance is still 53 %, which implies that it is low.

```
# best2 <- regsubsets(sum ~ . , data = bd, really.big=T)
# summary(best2)
# plot(best2, scale = "r")
```

It describes that Holiday (specially a Saturday), and the type of Western food are statistically significant to predict.

As a conclusion for this section, I have to state that no one model has a great fit, but the prediction is done with the month variable. As result the preditions for the next 6 months are very closed to the mean. This is a result of the low fit of the model.

```
m5 <- lm(suma ~ mes , data = bd)
summary(m5)

mes <- data.frame(mes = c(20170601,  20170701, 20170801, 20170901, 20171001, 20171101))
```

```
predict(m5, newdata = mes , interval = "prediction")
```

```
##        fit       lwr      upr
## 1 8.634487 -8.536691 25.80567
## 2 8.652944 -8.518493 25.82438
## 3 8.671401 -8.500305 25.84311
## 4 8.689858 -8.482127 25.86184
## 5 8.708316 -8.463959 25.88059
## 6 8.726773 -8.445800 25.89935
```

6. Based on the data and your ideas, plan strategies to double the total restaurant visitors in six months.

I would like to use 2 strategies:

- Based on the modelling techniques, I strongly use the variables that area statistical significant: Holiday, the type of Western food, the Restaurat, and the day of the week (specially a Saturday)
- Based on the professional expertise, I will hire a specialist in restaurant (because nowadays no one machine or statistical model is better that the human expertis to simulate or predict in long-term situations, specially in complex or chaotic environments, as selling or human interactions).
- Based on mi experience as a business model designer, designing statistical experiments of advertising promotions, in order to know what strategy and variable allow to have higher sales. Once the variables are known, design a sales model collaboratively with prototypical customers

7. **- Imagine that these restaurants are in your city (and not in Japan), what other data would you want to join in order of get more insights to increase the visitors?**

8. **How many channels can you think of downloading a DiDi Rides APP and how will you estimate the quality and cost of each channel?**

9. **We want to build up a model to predict "Possible Churn Users" for DiDi Rides APP (e.g.: no trips in the past 4 weeks). Please list all features that you can think about and the data mining or machine learning model or other methods you may use for this case**