DEPARTAMENTO
DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Trabajo Práctico 1

5 de septiembre de 2017 — Ciencia de datos

Integrante	LU	Correo electrónico
Christian Cuneo	755/13	chriscuneo93@gmail.com
Julián Bayardo	850/13	julian@bayardo.com.ar

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria - Pabellón I
Intendente Güiraldes 2160 - C1428EGA
Ciudad Autónoma de Buenos Aires - Argentina
Tel/Fax: (54 11) 4576-3359
http://exactas.uba.ar

Índice

1. Exploración Preliminar

Comenzamos graficando los datos para ver bien con que estamos trabajando.
Primero un simple gráfico de lineas:

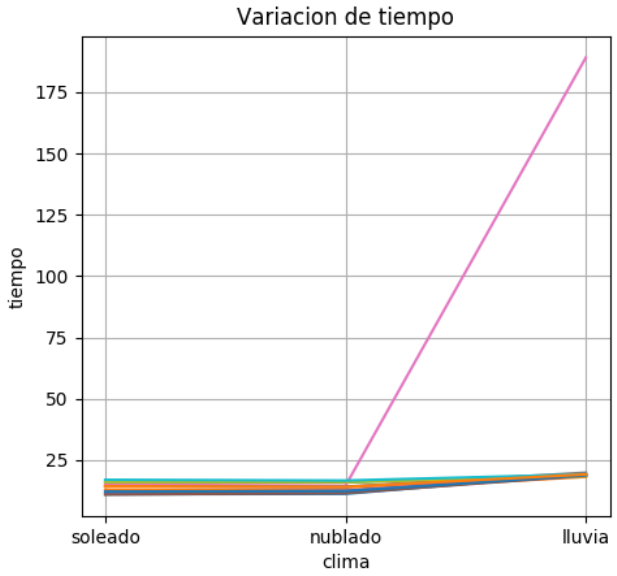

Figura 1: Vista general

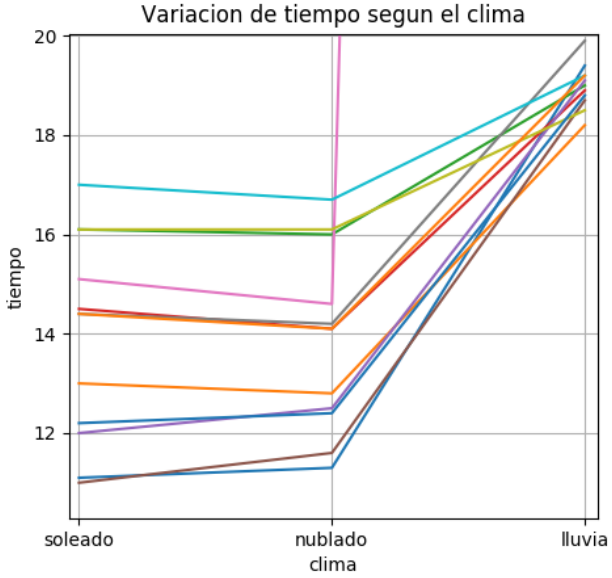

Figura 2: Zoom

Ya aquí podemos ver a simple vista que hay un dato que no tiene sentido, ya que estamos hablando de humanos corriendo, por lo tanto desde nuestro punto de vista no tiene sentido una diferencia tan amplia de tiempo.

A continuación procedimos a limpiar los datos, únicamente este tiempo en lluvia para el sujeto 7. La intuición inicial nos diria que directamente borremos toda esa fila, y demos como invalida las mediciones del sujeto 7, pero al tener tan pocos datos, decidimos solamente limpiarlo, suponiendo que lo que se olvidaron de poner es el punto decimal, trasnformando el 189 en un 18,9

A continuación veremos los datos limpios:

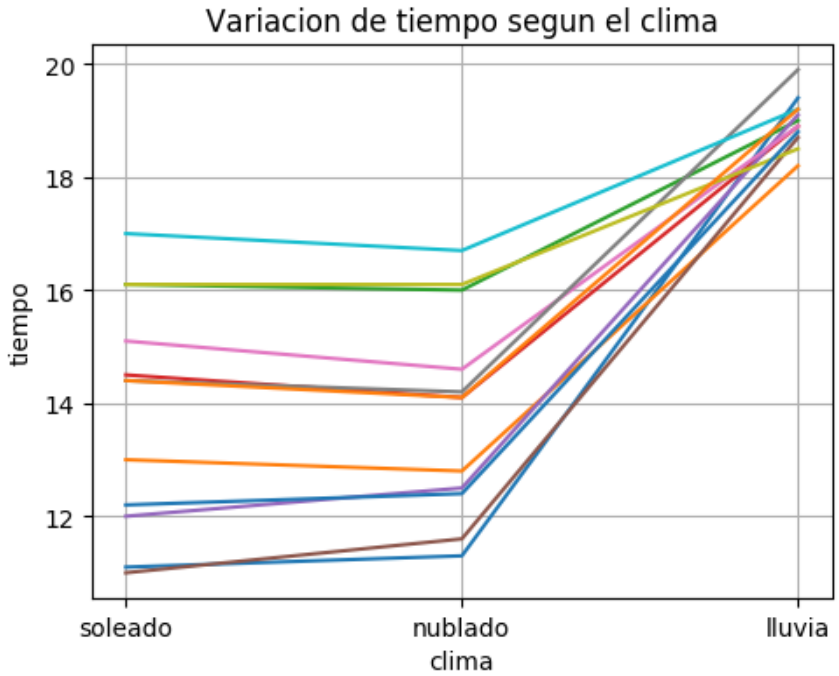

Figura 3: Datos limpios

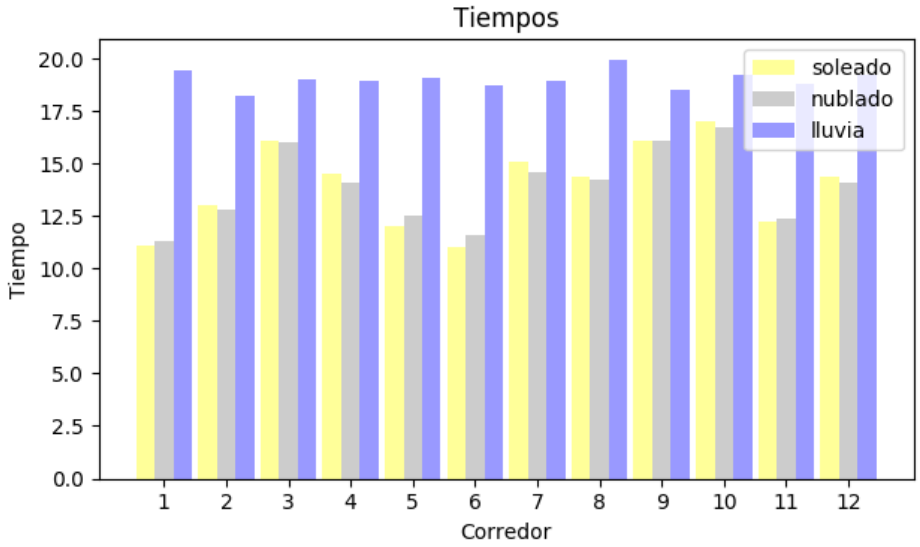

Figura 4: Grafico barras datos limpios

2. Tests

Ya con datos confiables procedemos a hacer pruebas estadísticas. Primero hicimos lo mas básico que fue plotear boxplots de los sampleos para cada clima:

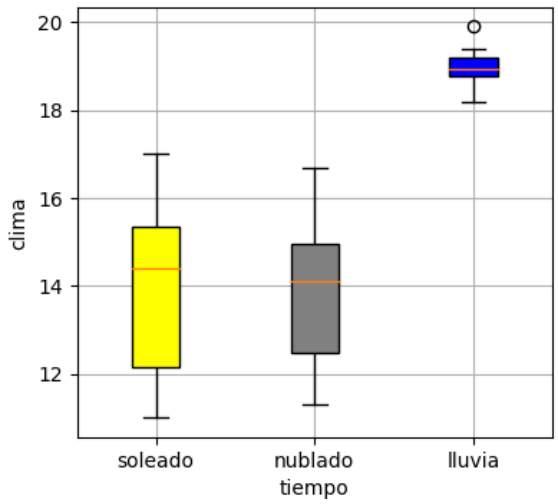

Figura 5: Media y desviación estándar según el clima

Ya podemos ver a simple vista que no hay diferencia perceptible entre la corrida con sol y la nublada, sino que la variación principal se puede ver al correr bajo lluvia. Acá podemos ver las diferencias entre las corridas, también en formato boxplot, la diferencia se calcula para cada sujeto:

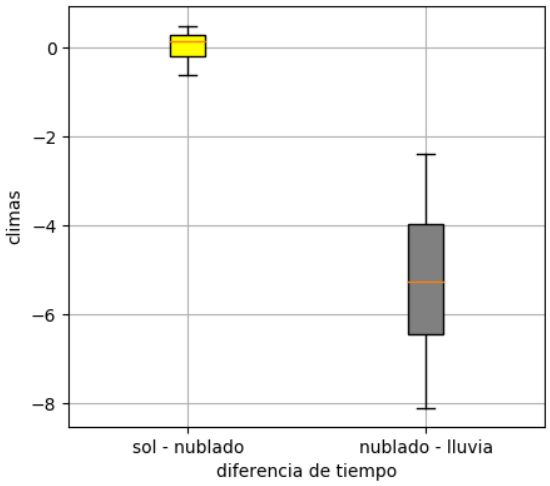

Figura 6: Media y desviación estándar de las diferencias

Vemos que la diferencia es insignificante entre el soleado y el nublado, pero no así entre el nublado y la lluvia (o el solado y la lluvia)

A continuación empezamos a realizar las pruebas, primero la mas común, un t-test para sampleos apareados, los resultados siendo los siguientes:

- Sol vs Nublado: 0.4121382498605874 con pvalue=0.6881556115604558
- Nublado vs Lluvia: -9.720128995157081 con pvalue=9.810399492877932e-07
- Sol vs Lluvia: -8.576227572594302 con pvalue=3.3516511607218045e-06

El resultado muestra (con pvalues que indican resultados significativos) que podemos descartar la hipótesis nula (que diría que no hay una diferencia entre las medias de los grupos) para los pares nublado vs lluvia y soleado vs lluvia.

Luego decidimos hacer un test no paramétrico, el signed rank de Wilcoxon:

- Sol vs Nublado: 28.5 con pvalue=0.6882065030738291
- Nublado vs Lluvia: 0.0 con pvalue=0.002217721464237049
- Sol vs Lluvia: 0.0 con pvalue=0.002217721464237049

Viendo de nuevo que la probabilidad de la hipotesis nula para los casos nublado vs lluvia y sol vs lluvia es rechazada, con un pvalue ¡0.01.

Luego, realizamos un test de permutaciones, con 1000 permutaciones distintas, solo para la diferencia entre soleado y lluvia, ya que a esta altura ya vemos que la diferencia relevante se encuentra ahí y obtuvimos que la probabilidad de la hipótesis nula es prácticamente 0:

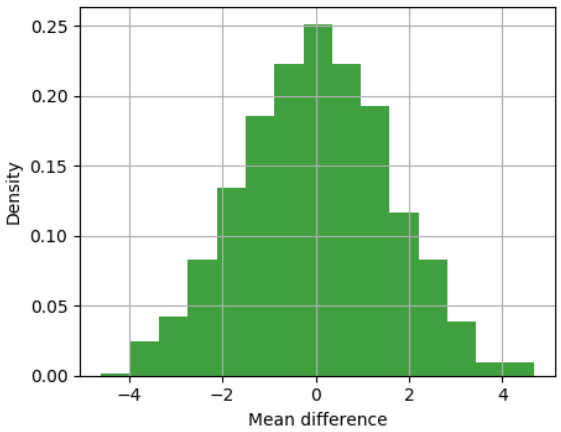

Figura 7: Histograma test de permutaciones

Por ultimo calculamos el coeficiente de correlación de pearson:

- Sol vs Nublado: 0.9908237767557433 con pvalue=5.045646399806607e-10
- Nublado vs Lluvia: 0.04228906602480422 con pvalue=0.896177007833118
- Sol vs Lluvia: 0.0525893206142193 con pvalue=0.8710570171791218

Aquí podemos ver que el único par que muestra cierta correlación es la del sol vs nublado, que vimos que no tenían una diferencia significativa. Aunque podemos ver que el pvalue para los otros dos pares muestra que el resultado no es significativo, por lo que no se puede asumir una correlación.

3. Conclusiones

1. Si, y podemos ver este resultado con todos los test estadísticos que corrimos (no el de pearson, porque no aplicaría a este caso). En todos vimos que la hipótesis nula (que las diferencias no son significativas) fue rechazada.

2. También lo confirmamos, viendo que la hipótesis nula es aceptada al comparar las corridas soleadas con las nubladas, en todos los tests.

3. A simple vista parecería que no, ya que es consistente la velocidad en los otros climas, con la velocidad con lluvia, solo que se alentece para todos los corredores. El coeficiente de pearson dio una correlación positiva, por lo que se mostraría que se comportan de igual manera, pero el pvalue nos dice que puede no ser significativo, aunque con tan pocas muestras no es tan confiable.

4. Esta conclusión no podemos sacarla de esta investigación ya que el clima puede ser muy variado y no analizamos todos los casos acá, ya que tenemos mediciones limitadas a 3 casos de clima.