Spyder es un entorno de desarrollo, un entorno de ejecución de Python (en local), no como Notebook (este es mejor para dar clase o para presentar un informe).

Si doy solo “a” lo invoco y si pongo “a” “b” me sacará solo el último, si quiero que me salga todo tengo que hacer print de eso.

Int (entero, variable numérica) y Str (texto, variable alfanumérica). Y también están las listas o los DataFrames, entre otros.

# al principio es para comentarios (para humanos).

#%% es para crear bloques de código independientes.

%reset –f es para borrar todo el entorno de ejecución (después poner una # al principio siempre mejor).

El play grande verde ejecuta todo el script, una a una es el negro con el stop.

Del es para borrar variables.

Listas es una sucesión de valores (sean del tipo que sean) separados por comas y enmarcadas entre corchetes [].

X= Pd.DataFrame({ para crear una tabla.

Los casos por filas (alumnos) y los atributos o características por columnas (edad, género, etc.). Es lo mismo que decir que la primera columna serán los casos (name, p.e.).

En csv americano las comas separan y el punto señala decimal; en csv europeo es punto-coma y coma, respectivamente. Mirarlo antes de importar datos.

La gestión de ficheros con pandas, las gráficas con matplotlib (librerías).

Plt.show() es para detener la gráfica.

Unic ID o clave (on) para lincar los dos DataFrames (rentals y weather) y hacer uno solo que agrupe los dos, que será la columna del day o la de fecha porque son comunes las dos, pero mejor la del día porque así no hay caracteres extraños como la barra en la fecha.

 Valuable forums:

* https://stackoverflow.com
* https://www.datacamp.com

 Valuable tutorials:

* https://datatofish.com/python-tutorials/
* https://pbpython.com/
* <https://matplotlib.org/tutorials/index.html>

﻿﻿Cuando se añaden casos hacia abajo con el append, el index se corta y empieza de nuevo con el nuevo DataFrame. No tiene que tener en cuenta la variable (columna) index al importar, tiene que llegar el index hasta 730. Se usa ignore\_index=True para esto.

Hay variables nominales (estaciones, día de la semana, situación del tiempo) y cuantitativas (temperatura, número de alquileres) y cada una se analizará de una manera en función de su tipo, con métodos estadísticos distintos.

Media y desviación típica son los dos descriptivos básicos de una variable **cuantitativa**. Y gráficamente la describimos con un histograma.

De una variable **cualitativa** miraremos porcentajes y gráfico de barras, de frecuencias relativas o porcentajes (‘barchart’).

* Descargar un dataset de Kaggle y hacer análisis descriptivos con 6 o 7 variables. Hacer un informe en notebooks de Colab por ejemplo que apunte a GitHub. Mirar, cargar los datos, limpiarlos y analizar las variables.

Muchas veces hay valores de 96,97,98 o 99 para reportar que es un valor no válido, en una encuesta, por ejemplo, cuando no hay respuesta. No son valores posibles de temperatura, no había datos esos días y asignó esos valores. Hay que tener cuidado y reemplazarlos por nan. Pero…

Pero… Si en el DataFrame hay huecos (nan) los histogramas u otros gráficos puede que no funcionen, por lo que hay que eliminarlos (.dropna).

Siempre que recodificas es mejor hacerlo sobre una nueva variable, no cargarte la previa, por si acaso. Por tanto, nunca machacar las variables. ¡Si hay una recodificación o transformación se debe hacer en una columna nueva!

Para clasificar por rangos y agrupar los casos, se puede utilizar la desviación típica (los puntos a una desviación) o los cuartiles (estos también aparecen en los descriptivos: 25% y 75%).

Gráficos: https://aprendeconalf.es/docencia/python/manual/pandas/

Una variable cuantitativa contra una nominal = **comparación de medias**. Dos pruebas: la prueba de t (2) y la ANOVA (>2). Si quiero saber si vendo más entre semana o el fin de semana, pues hago una media de cada tipo de día (2 grupos) y comparo. Tipo de día (cualitativa) y ventas (cuantitativa): quiero ver si ‘working day’ tiene un impacto sobre las ventas. Para ello, contraste de hipótesis. La hipótesis nula (Ho) parte de que **NO** hay asociación entre las variables que estamos estudiando. Para irnos a la alternativa (H1), para asegurar que sí hay asociación entre la variable cualitativa y la cuantitativa, tenemos que tener una confianza alta (al menos del 95%).

Comparación de medias entre dos grupos: t-test para muestras independientes y gráfico de intervalos de confianza (plot the means).

Para saber si mi muestra es representativa de la población en cuanto a generalizar mi hallazgo de asociación entre variables no solo veremos las medias, también las varianzas o desviaciones, así como el grado de confianza (95% o más).

P-value es la probabilidad de equivocarnos al rechazar la hipótesis nula (Ho). Buscamos un valor cuanto más pequeño mejor (al menos menor de 0.05 o un 5%). Es decir, confianza de 95%. Y suponemos que la muestra es aleatoria e independiente, que la muestra es buena vamos. No nos preocupamos por esto a este nivel.

La prueba de t es para saber si podemos extrapolar lo deducido en la muestra para la población en general o si solo sirve para los datos de la muestra analizados y no nos sirve entonces para la población.

El margen de error de la media de uno y otro al 95% se solapa, no nos atrevemos a aseverar nada, por eso p-value daba alto: con uno del 11% no podemos afirmar que haya relación. Gráfico de intervalos de confianza. La media global es la línea verde. El intervalo de confianza es el margen de error y es simétrico respecto a su media.

Se **reporta** el estadístico t (es el t-student), el p-value y la n. Aparte si queremos la media global también. Lo ideal es combinar estos datos numéricos junto con el gráfico asociado en el mismo dibujo.

Primer paso, mirar medias de la muestra; segundo, intento extrapolar a la población, pero si el p-value es tan alto no me atrevo a rechazar la Ho (no hay relación). Me quedo con la idea de que los alquileres son iguales independientemente del tipo de día, no hay diferencias significativas, son compatibles con el error de muestreo y puede ser que esas diferencias en la muestra se deban al azar. La media poblacional podría ser diferente a la media muestral y no ser cierto que una es mayor a la otra y realmente sea al contrario.

En general, si subo la confianza (del 95% al 99%, por ejemplo), los márgenes de error o intervalos de confianza aumentarán, obviamente. Y habrá más riesgo de que se solapen.

Si hay más de dos grupos entonces sería test **ANOVA** de varianzas, en vez de t-test. Mide, por ejemplo, si hay diferencias en función de la situación del tiempo al 99% de confianza entre los 3 grupos. ANOVA te dice que no hay probabilidad prácticamente de que los 3 grupos sean iguales, pueden ser los 3 diferentes o solo 2 de ellos. Para saber cuáles exactamente, hacemos el gráfico. Aquí los márgenes de error o intervalos no se superponen entre los diferentes grupos, el p valor es pequeño, entonces nos vamos a la hipótesis alternativa de que al menos dos de ellos difieren en cuanto a las medias de alquileres, no son iguales los 3 grupos (rechazamos Ho). Por tanto, concluimos que en términos poblacionales las diferentes condiciones del tiempo afectan significativamente a la media de alquileres diarios.

¡¡Siempre describir primero las variables!!

Las variables nominales o cualitativas las describimos con porcentajes. Y si comparamos dos nominales haremos una **comparación de porcentajes** para ver si hay asociación entre ellas. Tablas de contingencia o cross tabulation: es una tabla de doble entrada donde tienes la frecuencia de cada una. Esto para describirlas. Luego, para extrapolarlo a la población haremos la prueba de Chi-2. Y como opcional la V de Cramer. Y aparte, están los gráficos (de barras agrupados).

Hoy haremos que la variable dependiente (‘cnt’) sea cualitativa, el otro día hicimos la comparación de medias cuando era cuantitativa. La variable dependiente es alquileres y la independiente es el tipo de día. La hipótesis nula es que los alquileres son independientes del tipo de día.

Utilizamos como referencia la distribución porcentual de la variable dependiente (la columna de total) para comparar cuando tenga las distribuciones de los grupos. Hago normalización **por columnas**, porque me interesa saber dentro de un tipo, cuántos son altos, medianos y bajos. Tienen que sumar 100 por COLUMNAS.

Los días buenos pasan más en los laborables y los malos pasan más en los festivos. ¡¡¡Pero en la muestra!!! Necesito hacer un test estadístico para ver si puedo generalizar para la población: Chi2.

Para el estadístico Chi2 se coge la ct original, solo con las dos variables, sin total y sin porcentajes, ¡solo frecuencias! El alto p-value nos dice que a nivel poblacional las ventas no dependen del tipo de día (‘wd’).

Gráficamente hay que asegurarse de que el total es el mismo que el total que ya habíamos dibujado al principio, al describir la variable cualitativa dependiente. Y dibujamos lo mismo para los n grupos.

**Reportar**: la tabla de contingencia de doble entrada (con el estadístico, n y p-value al pie) y el gráfico de barras combinado (con el estadístico, n y p-value en cuadro de texto).

La ‘V de Cramer’ nos dice la fuerza de la asociación, cuando haces varios tests entre variables nominales y todos tienen p-valor muy bajo, por ejemplo. Para saber cuál tiene mayor influencia. Y los gráficos de mosaico son muy bonitos también cuando se hacen varios análisis de este tipo.

Colores en matplotlib:

https://matplotlib.org/2.0.2/examples/color/colormaps\_reference.html

Relación entre variables cuantitativas: **correlación y scatter plot**

El coeficiente de correlación de Pearson mide la relación lineal entre dos variables cuantitativas. La hipótesis nula es que r=0, no hay correlación, no hay asociación entre variables: eso significa que en el gráfico hay una nube de puntos idénticamente y uniformemente distribuidos, sin ninguna zona con mayor densidad de puntos que otra.

El coeficiente puede ir [-1,1], pasando por 0 cuando no hay asociación absoluta. Signo negativo es relación inversa y signo positivo es relación directa. Lo vendemos con un cuadrito de texto con n, r y p-value. Y si podemos ponerlo junto en un panel los dos gráficos queda mejor.

Vemos que hay cierta no-linealidad (relación curva al final, cuando sube mucho la temperatura bajan las ventas de bicis). Y también observamos que al principio hay menos dispersión y luego más (se abre): ello se podría explicar porque la demanda es elástica en función de la temperatura. Hay dos pendientes. Vemos “dos delfines” que podrían ser los dos años (ya que n=731).

*Nota*: buscar ‘markers’ para Python en Google, para los gráficos: poner cuadrados, p.e.

El p-value mide la probabilidad de equivocarme cuando digo que r!=0.

La parte no lineal el coeficiente de Pearson no la captura: si hay relación de parábola saldrá que r=0 porque una cancela la otra, ¡pero sí hay relación! Lo que pasa es que no es lineal. Podría dividir y hacer sub-setting en dos, entonces me saldría una r alta positiva y una r alta negativa. Partiré en dos en el punto donde las dos r sean máximas (punto de inflexión).

¡Correlación NO es causación! Causación requiere correlación, pero no a la inversa.

**REGRESIÓN**

Siempre describir antes las variables, antes de meterlas en el modelo. Vamos a aprender con dos variables cuantitativas. Por tanto, antes también se deberá explorar su correlación, tal y como aprendimos.

Para definir una recta necesito saber el punto de partida (intercepto o constante o ‘a’) y el ángulo (pendiente o beta). Me permite predecir las ventas en función de la temperatura que haga. De momento solo tengo un predictor. OLS: es modelo ordinario de mínimos cuadrados. De lo que nos da nos interesa la pendiente (coeficiente del predictor) y el intercepto, así como sus p-values, los cuales son pequeñísimos: son significativos, me creo que en la población funcionará como en mi muestra.

Aparte está n (número de observaciones) y el coeficiente de determinación (R cuadrado), el cual varía de 0-1. El 0 es que el predictor no pinta nada para predecir la ‘y’; y el 1 es justo lo contrario, que el modelo es plenamente determinista. R2 es la cantidad de variabilidad que hay en la variable dependiente (ventas) que puedo asociar o que se explica por la variabilidad de las variables independientes del modelo (temperatura). También se habla en términos de *varianza explicada* por la temperatura. Conociendo la temperatura, puedo explicar el 40% de la variabilidad de las ventas.

Hay que buscar la recta que minimiza la suma de los errores o residuos (las distancias de los valores reales a los predichos por el modelo); realmente sus cuadrados, para evitar valores negativos (modelo OLS).

*Nota*: Cuando metes predictores que no explican bien (R2 muy pequeña) es mejor no ponerlos en el modelo (*overfitting* del modelo).

*Nota*: Slope es pendiente.

Si no hubiera asociación alguna, habría una nube de puntos uniformemente distribuida sobre el plano real, y por tanto la pendiente de una hipotética recta de regresión tendría pendiente nula (b=0), sería una recta horizontal que cortaría el eje ‘y’ en la media.

**Modelo de regresión múltiple:** tenemos varios predictores.

Ventas= 1991 + (156\*t) – (52\*ws) + ë (error o épsilon porque R2 sigue siendo < 1). Cuánto se incrementa la variable dependiente cada vez que aumenta una unidad la variable independiente: coeficiente de cada predictor; eso sí, *ceteris paribus* (mientras no cambien el resto de predictores del modelo múltiple). Cambian los coeficientes respecto a cuando era regresión simple, porque uno explicaba parte del otro (como en el ejemplo de color de piel y renta sobre el CI). Es decir, estimamos el impacto de cada variable independiente cuando controlamos el impacto del resto de variables del modelo (se mantienen constantes). Cuando hacíamos regresión simple (1 predictor) era sin controlar al resto, por eso cambian los coeficientes. Cuando meto un nuevo predictor mejor hacerlo simple primero y luego ya meterlo en el modelo múltiple.

Por tanto, la regresión permite CONTROL estadístico, y cuantos más controles haya (más predictores), más certeza tienes en el modelo. Es decir, el coeficiente de un predictor muestra la asociación con la dependiente manteniendo el resto de variables constantes, controlándolas.

(Stargazer) Poner en Google: html render online para visualizar html.

El intercepto es el valor de la predicción cuando el resto de las variables valen 0.

Variables **dummies**, son falsas, son para capturar la información que tenía en otra, a fin de convertirla en una dicotómica de 0 y 1, porque con la original no podía, no tenía sentido (por ejemplo, en weekday o en weathersit). Son cualitativas.

Un diccionario es una tabla de equivalencias.

Si introducimos una variable dummy hay que sacar una fuera del modelo que sirva como categoría de referencia. Si saco sunny, el intercepto mide las ventas cuando esa es 1. Lo normal es dejar fuera la más común.

Metemos tanto la variable lineal como su cuadrado, porque al final baja. Para linealizar el modelo, dado que vemos en el scatter plot que hay una relación cuadrática con la temperatura.

Logistic Regression statsmodels: sm.Logit; sino pues regresión lineal (si hay dos).

Regresión logística:

Es significativa la variable temperatura (p-value), es decir, importa; y es positivo el coeficiente (sigma). También haremos caso al pseudo-R2 (que es bajo en este caso). Se podría medir también la probabilidad, pero no lo vamos a ver a este nivel.

Regresión logística multi-nomial es para cuando hay más de dos opciones (no una dicotómica simplemente), donde se lanzan varias regresiones logísticas a la vez. Aunque también podría hacerse a nuestro nivel enfrentando a una contra una, n veces, hasta realizar todas las comparaciones posibles. Ejemplo de elegir marca de coche que comprar, dependiendo del salario o de otras variables en el lado de la demanda.

Cuidado con las técnicas inductivas y con meter todo en un modelo; hay que medir bien qué meter en el modelo. False Discovery Rate (FDR) en la comparación múltiple de valores aleatorios, donde se encuentran p-valores significativos en relación a las correlaciones entre los mismos.