

Contenido

Descripción del dataset	2
Limpieza de los datos	2
Selección de los datos de interés a analizar	2
Limpieza de datos.....	3
Comprobación de campos – Embarcados y Fallecidos concordantes	3
Comprobación de campos – Date válido	3
Comprobación de campos – Operator válido	4
Comprobación de campos – Embarcados y Fallecidos rellenos	4
Comprobación de campos – Eliminación de los datos de aviación militar.	5
Comprobación de campos – Valores extremos.	5
Comprobación de campos – Años incompletos.....	5
Análisis de los datos.	6
Selección de los grupos de datos que se quieren analizar/comparar.	6
Comprobación de la normalidad.....	7
Comprobación de homogeneidad de la varianza.....	7
Aplicación de pruebas estadísticas	8
Representación de los resultados a partir de tablas y gráficas.....	9
Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones?	11
Recursos	12

Autor: Rodrigo Minguito Linaje

Descripción del dataset

El dataset es Airplane_Crashes_and_Fatalities_Since_1908.csv, disponible en www.kraggle.com.

Con el análisis de este dataset se pretende verificar si los accidentes que se producen en aviación civil causan cada vez menos fallecimientos.

Insisto: lo que pretendo es verificar si los accidentes en aviación comercial son cada vez más “leves”. De entrada no busco verificar que cada vez hay menos accidentes, ni que el ratio de accidentes por número de vuelos en el año es menor.

Limpieza de los datos

Selección de los datos de interés a analizar

¿Qué datos me interesan?

Las columnas que suministra el dataset son:

- Date: Fecha
- Time: Hora
- Location: Lugar
- Operator: Aerolínea
- Flight #: Número de vuelo
- Route: Ruta
- Type: Tipo de Avión
- Registration: Matrícula del avión
- cn/In
- Aboard: Personas embarcadas
- Fatalities: Personas fallecidas
- Ground: ¿el accidente tuvo lugar en tierra?
- Summary: resumen

De ellos, para el análisis a realizar, me quedaré con:

- Obviamente, la fecha del accidente, sin necesidad de la hora, para poder filtrar en un lado u otro.
- El tipo de vuelo, si era militar o civil, para lo que necesitaré filtrar la aerolínea, eliminando todas las que entradas que contengan ‘Military’ o ‘Ejército’
- El número de personas embarcadas¹
- El número de personas fallecidas

¹ Al final no lo utilizaré para nada

Limpieza de datos

¿Los datos contienen ceros o elementos vacíos? ¿Y valores extremos? ¿Cómo gestionarías cada uno de estos casos?

Para empezar, antes de codificar nada, tomo el fichero y me quedo con los datos que me interesan, esto es:

- Elimino todas las filas anteriores a 1950
- Elimino todas las columnas que no me interesan, para quedarme solo con Date (A), Operator (B), Aboard (C) y Fatalities (D)

Con estos datos, realizo una serie de comprobaciones.

Comprobación de campos – Embarcados y Fallecidos concordantes

En Excel, a partir de esos datos, añado una nueva columna, con la fórmula `=SI(C2<D2;1;0)` a todas las filas con datos. Y calculo la suma total... 6 elementos que tienen más fallecidos que embarcados ¿¿¿qué???

	A	B	C	D	E	F	G	H
1	Date	Operator	Aboard	Fatalities	Ground			
440	04/20/1957	Air France		1	0	1		
1968	11/03/1977	El Al		1	0	1		
2284	12/16/1981	Bristow Helicopters		12	0	1		
2330	08/11/1982	Pan American World Airw		1	0	1		
2804	05/09/1989	Aero Asahi		10	0	1		
3041	02/20/1992	Aerolineas Argentinas		1	0	1		

He recuperado, para este caso, el concepto Ground, por si se trataba de aviones que se habían estrellado causando muertos en tierra, pero no parece el caso.

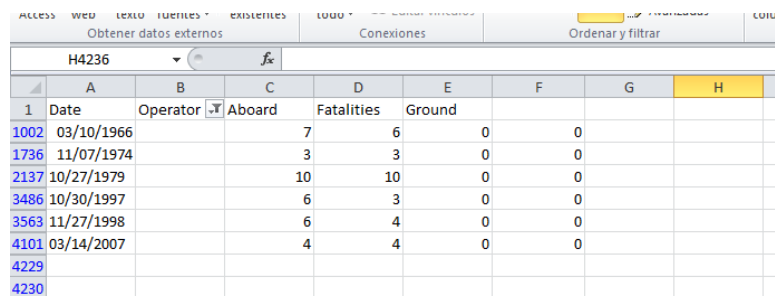
Por tanto, para estos 6 casos tomo la decisión de rellenar el campo Aboard con el mismo número que el de Fatalities

Comprobación de campos – Date válido

En este caso podría hacer comprobaciones de que realmente todas las fechas concuerdan con el formato mm/dd/aaaa y eliminar las no válidas.... Pero lo que me interesa es el año, no la fecha completa, por tanto, en la carga de los datos, podré hacer un tratamiento directo de cadenas, tomando el tercer token, cortando por el carácter '/'

Comprobación de campos – Operator válido

Desde Excel, aplico un filtro sobre la columna Operator, y me quedo sólo con las entradas vacías.

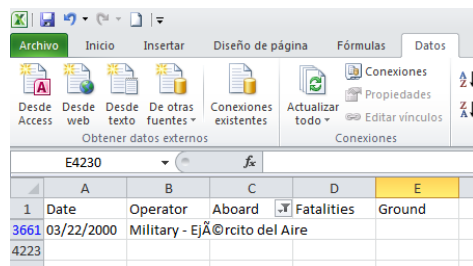


	A	B	C	D	E	F	G	H
1	Date	Operator	Aboard	Fatalities	Ground			
1002	03/10/1966		7	6	0	0		
1736	11/07/1974		3	3	0	0		
2137	10/27/1979		10	10	0	0		
3486	10/30/1997		6	3	0	0		
3563	11/27/1998		6	4	0	0		
4101	03/14/2007		4	4	0	0		
4229								
4230								

En este caso, como no puedo discernir si se trata de registros civiles o militares, elimino las entradas afectadas

Comprobación de campos – Embarcados y Fallecidos rellenos

Para este caso, aplico filtros a las dos columnas, y selecciono los valores <Vacio>

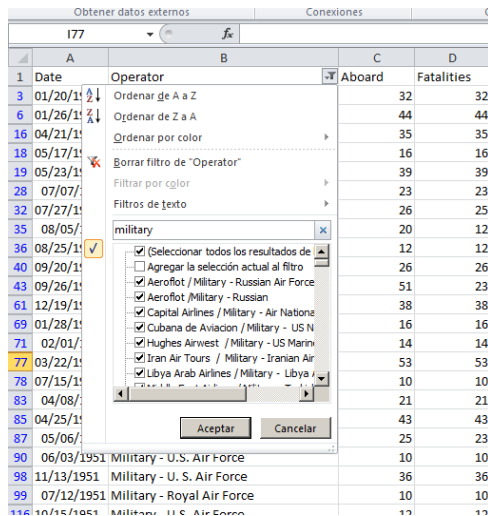


	A	B	C	D	E
1	Date	Operator	Aboard	Fatalities	Ground
3661	03/22/2000	Military - Ejército del Aire			
4223					

Da como resultado un solo registro, que elimino.

Comprobación de campos – Eliminación de los datos de aviación militar.

Este tratamiento lo podría hacer desde el código de Python, pero ya que tengo las funciones de Excel, puedo verlo de forma más visual y ahorrar procesamiento.



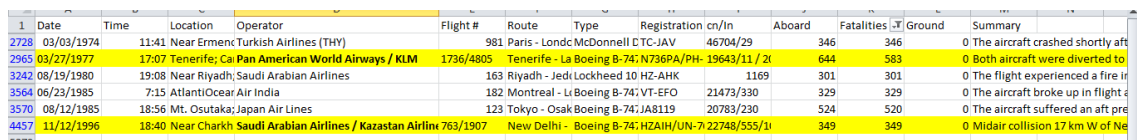
	A	B	C	D
	Date	Operator	Aboard	Fatalities
1				
3	01/20/11	Ordenar de A a Z	32	32
6	01/26/11	Ordenar de Z a A	44	44
16	04/21/11	Ordenar por color	35	35
18	05/17/11	Borrar filtro de "Operator"	16	16
19	05/23/11	Filtrar por color	39	39
28	07/07/11	Filtros de texto	23	23
32	07/27/11		26	25
35	08/05/11	military	20	12
36	08/25/11		12	12
40	09/20/11		26	26
43	09/26/11		51	23
61	12/19/11		38	38
69	01/28/11		16	16
71	02/01/11		14	14
77	03/22/11		53	53
78	07/15/11		10	10
83	04/08/11		21	21
85	04/25/11		43	43
87	05/06/11		25	23
90	06/03/1951	Military - U.S. Air Force	10	10
98	11/13/1951	Military - U. S. Air Force	36	36
99	07/12/1951	Military - Royal Air Force	10	10
116	10/15/1951	Military - U.S. Air Force	12	12

Por tanto, meto un filtro por la columna Operator y busco los registros *Military* y *Ejército*

Comprobación de campos – Valores extremos.

¿Qué sería en este caso un valor extremo? Considero que podría ser cualquier caso donde haya más Fatalities que la media del tamaño medio de un avión. Supongamos que el tamaño medio es de 300 plazas.

Aplico un filtro por ese valor y obtengo seis registros. Ahora será cuestión de comprobar a ojo su descripción para ver si tiene sentido:



	Date	Time	Location	Operator	Flight #	Route	Type	Registration	cn/in	Aboard	Fatalities	Ground	Summary
2728	03/03/1974	11:41	Near Erment	Turkish Airlines (THY)		Paris - Londc	McDonnell DTC-JAV	46704/29		346	346		0 The aircraft crashed shortly aft
2965	03/27/1977	17:07	Tenerife; Ca	Pan American World Airways / KLM	1736/4805	Tenerife - La	Boeing B-747-N736PA/PH-19643/11/21			644	583		0 Both aircraft were diverted to
3242	08/19/1980	19:08	Near Riyadh;	Saudi Arabian Airlines		163 Riyadh - Jed	Lockheed 10 HZ-AHK		1169	301	301		0 The flight experienced a fire in
3564	06/23/1985	7:15	AtlantiOcear	Air India		182 Montreal - L	Boeing B-747-VT-EFO	21473/330		329	329		0 The aircraft broke up in flight a
3570	08/12/1985	18:56	Mt. Oosaka;	Japan Air Lines		123 Tokyo - Osak	Boeing B-747-JA8119	20783/230		524	520		0 The aircraft suffered an aft pre
4457	11/12/1996	18:40	Near Charkh	Saudi Arabian Airlines / Kazastan Airlin	763/1907	New Delhi -	Boeing B-747-HZAIH/UN-7122748/555/11			349	349		0 Midair collision 17 km W of Ne

En dos de los registros intervienen dos aviones, por lo tanto, pueden tener sentido.

Para el caso más abultado (520), una búsqueda en la Wikipedia confirma que se produjo dicho accidente.

Para el caso de 329 fatalities, el tipo de avión (Boeing B747) es el mismo del caso anterior, por lo que los números pueden ser válidos.

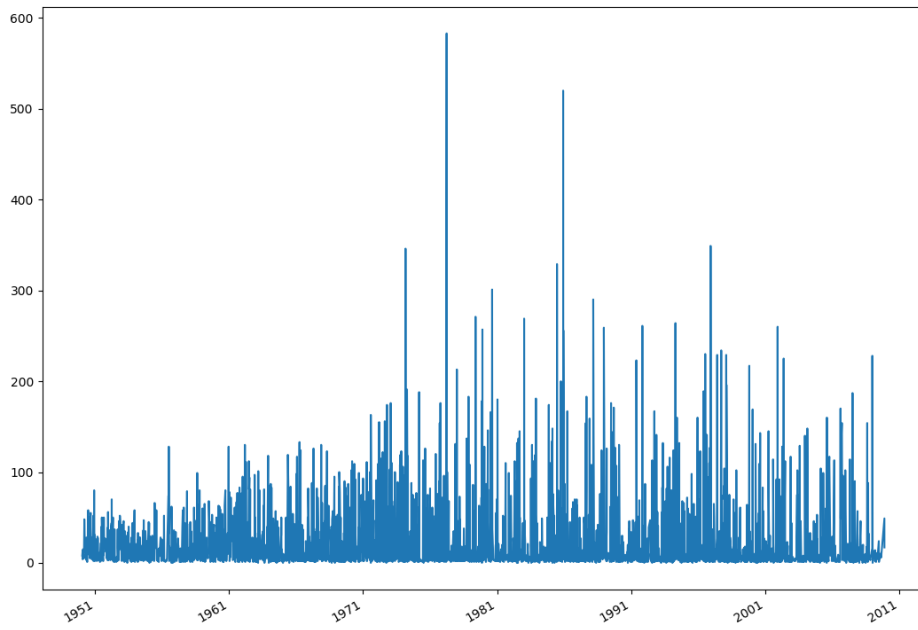
Para los dos registros restantes, las búsquedas en la Wikipedia confirman que son correctos.

Comprobación de campos – Años incompletos.

En este caso no figura ningún accidente posterior al 07/06/2009, lo que hace suponer que el año no está completo, por lo que elimino las entradas existentes para no contaminar el resto de años

Análisis de los datos.

Selección de los grupos de datos que se quieren analizar/comparar.



La representación de los datos de *Fatalities* de cada año, en crudo, queda como el gráfico superior. Pero esta representación es inútil, ya que los accidentes puntuales desvirtúan la representación, y además hay años que tienen muchos más accidentes que otros.

Por tanto, tomaré la media de las *Fatalities* de cada año.

Comprobación de la normalidad

Para verificar la hipótesis de la normalidad de los datos he usado varios métodos (<https://plot.ly/python/normality-test/>):

- Test D'Agostino and Pearson: El valor p es bastante cercano a 0 y mucho menor que el valor estadístico, por lo que no podemos rechazar la hipótesis.
- Criterio Kolmogorov-Smirnov: Dado que el valor p es 0, tenemos una fuerte evidencia para no rechazar la hipótesis.
- Criterio Shapiro-Wilk: Dado que el valor p es mucho menor que el valor estadístico, no podemos rechazar la hipótesis.

```
*****
Test de Normalidad
*****
Valor Estadístico 8.513952
Valor p 0.014165

*****
Test de Normalidad de Shapiro-Wilk
*****
Valor Estadístico 0.957275
Valor p 0.073176

*****
Test de Normalidad de Kolmogorov-Smirnov
*****
Valor Estadístico 1.000000
Valor p 0.000000
```

Para todos los casos cumplimos las condiciones que no se puede rechazar la hipótesis de que los datos procedan de una distribución normal.

Comprobación de homogeneidad de la varianza

Para esta comprobación uso los métodos de Lavene y Bartlett. Ejecuto cada uno de ellos 3 veces, y cada una de las pruebas usa 4 poblaciones de 10 elementos tomados al azar para hacer los cálculos.

```
*****
Homogeneidad de la varianza
*****
Levene Valor Estadístico 1.113106
Levene Valor p 0.356520
Levene Valor Estadístico 2.446481
Levene Valor p 0.079606
Levene Valor Estadístico 1.128040
Levene Valor p 0.350654

Bartlett Valor Estadístico 6.384819
Bartlett Valor p 0.094317
Bartlett Valor Estadístico 1.020920
Bartlett Valor p 0.796190
Bartlett Valor Estadístico 1.304629
Bartlett Valor p 0.728034
```

En todos los casos el valor p es mayor que 0.05, por lo que puede considerarse que las varianzas son homogéneas.

Aplicación de pruebas estadísticas

Lo que quiero es comprobar si existe una correlación lineal entre el año y el número de *Fatalities*, decreciente para la hipótesis de este caso.

Para ello correlaciono los datos *Anno* y *Media*.

```
*****  
Covarianza  
*****  
Covarianza -16.798456
```

La covarianza, por sí misma, no nos dice nada. En este caso, al ser negativo, nos dice, como mucho, que se relacionan de forma inversa.

Ahora trataremos de buscar la función (recta) que mejor se ajuste a todos los puntos, y determinar *cuánto* se ajusta. Ese *cuánto* se puede determinar sin necesidad de calcular la recta en sí, y se denomina Coeficiente de Correlación de Pearson. ¿Y qué indica ese valor?

```
*****  
Coeficiente de Correlacion de Pearson  
*****  
Coeficiente de Correlacion de Pearson -0.229357
```

En primer lugar, al ser negativo, como ya había visto con la covarianza, indica que existe una correlación negativa, lo que es lo mismo, a más año menos *Falalties*.

Pero en segundo lugar, el Coeficiente de Correlación de Pearson obtenido es muy bajo, lo que indica que apenas existe una relación lineal moderada entre ambas variables.

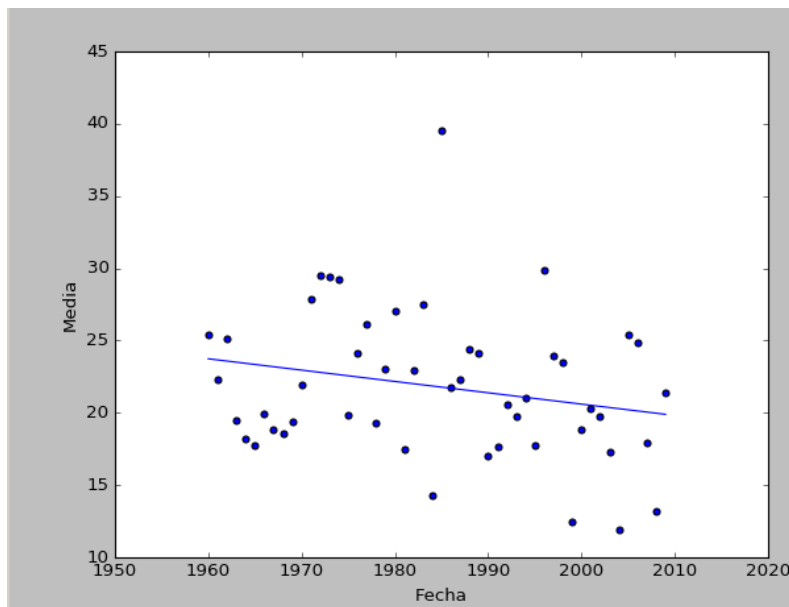
Representación de los resultados a partir de tablas y gráficas.

Los datos, con sus medias, quedan registrados en una tabla en la ejecución de Python:

	Anno	Acumulado	NumVuelos	Media
0	1960.0	1294.0	51.0	25.372549
1	1961.0	2300.0	103.0	22.330097
2	1962.0	2812.0	112.0	25.107143
3	1963.0	1927.0	99.0	19.464646
4	1964.0	1911.0	105.0	18.200000
5	1965.0	1793.0	101.0	17.752475
6	1966.0	2109.0	106.0	19.896226
7	1967.0	2376.0	126.0	18.857143
8	1968.0	2229.0	120.0	18.575000
9	1969.0	2408.0	124.0	19.419355
10	1970.0	2695.0	123.0	21.910569
...				
37	1997.0	1365.0	57.0	23.947368
38	1998.0	1434.0	61.0	23.508197
39	1999.0	871.0	70.0	12.442857
40	2000.0	1244.0	66.0	18.848485
41	2001.0	1279.0	63.0	20.301587
42	2002.0	1222.0	62.0	19.709677
43	2003.0	915.0	53.0	17.264151
44	2004.0	668.0	56.0	11.928571
45	2005.0	1116.0	44.0	25.363636
46	2006.0	920.0	37.0	24.864865
47	2007.0	861.0	48.0	17.937500
48	2008.0	727.0	55.0	13.218182

Media: 21.825580
Mediana: 21.025974
Varianza: 26.274223
Desviación estándar: 5.125839

Estos datos, representados gráficamente quedan como:



Como puede observarse, los puntos apenas 'aciertan' en la recta, lo que confirma visualmente lo calculado con el coeficiente de Pearson

Por curiosidad, repetiré los mismos análisis estadísticos, para verificar si cada año se producen menos *Fatalities*. En este caso:

```
*****
Media: 21.825580
Mediana: 21.025974
Varianza: 26.274223
Desviación estándar: 5.125839
*****

*****
Test de Normalidad
*****
Valor Estadístico 3.872617
Valor p 0.144235

*****
Test de Normalidad de Shapiro-Wilk
*****
Valor Estadístico 0.961658
Valor p 0.110855

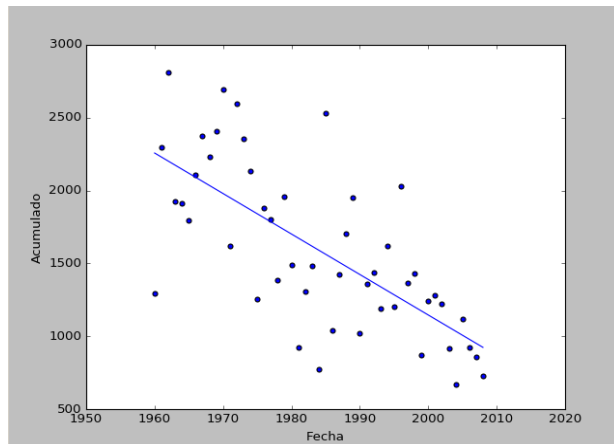
*****
Test de Normalidad de Kolmogorov-Smirnov
*****
Valor Estadístico 1.000000
Valor p 0.000000

*****
Homogeneidad de la varianza
*****
Levene Valor Estadístico 0.947233
Levene Valor p 0.428058
Levene Valor Estadístico 0.244966
Levene Valor p 0.864357
Levene Valor Estadístico 1.698819
Levene Valor p 0.184581

Bartlett Valor Estadístico 2.888946
Bartlett Valor p 0.409066
Bartlett Valor Estadístico 2.143446
Bartlett Valor p 0.543173
Bartlett Valor Estadístico 6.642354
Bartlett Valor p 0.084214

*****
Covarianza
*****
Covarianza -5674.520833

*****
Coeficiente de Correlacion de Pearson
*****
Coeficiente de Correlacion de Pearson -0.698630
```



En este caso se cumplen los criterios de normalidad y varianza, y presenta una correlación (Pearson) mucho más marcada.

Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones?

¿Qué puedo concluir?

En general, he tenido dos hipótesis:

- A cada año que pasa, ¿los accidentes que pasan son más pequeños?
- A cada año que pasa, ¿el número de fallecidos en accidentes es menor?

En una primera lectura podría pensarse que se trata de la misma hipótesis pero no es así. La primera podría replantearse como

*Si estoy en un accidente (con fallecidos) en un vuelo de aviación civil, ¿me conviene que sea en 1960 o en 2006?*²

Mientras que la segunda podría traducirse como:

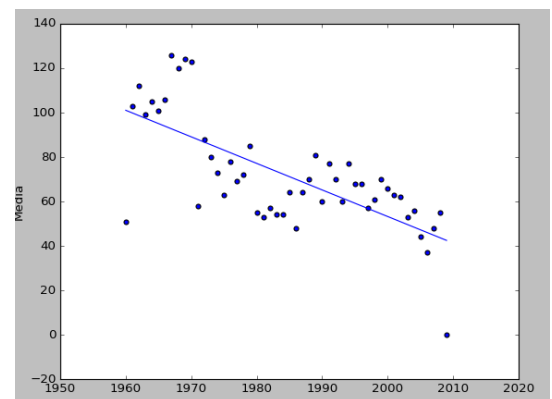
Las mejoras técnicas y las nuevas normativas, ¿están haciendo que los fallecidos en los accidentes de aviación civil sean cada vez menos?

La primera hipótesis ha quedado descartada viendo el coeficiente de Pearson, mientras que la segunda ha mostrado una fuerte correlación... Pero, ¿cómo puede ser esto, si los datos de la primera hipótesis salen del mismo sitio que los de la segunda? Pues porque hay un tercer jugador que he dejado deliberadamente escondido: el número de accidentes en total.

Con un análisis rápido (disponible de forma completa en el código), se obtiene una recta con un coeficiente de -0.703605 , lo que implica una gran correlación. Esto es: cada año hay menos accidentes.

En la primera hipótesis se está calculando una media, usando el número de fallecidos en el año (que como demuestra la segunda hipótesis está bajando) y el número de accidentes (que acabo de demostrar que también está bajando).

Al estar usando dos variables que están bajando para calcular una tercera, parece que las bajadas se compensan y acaban causando que no se obtenga correlación.



Como conclusión, queda demostrado que cada año se producen menos accidentes y que el total de fallecidos en los accidentes está bajando, pero no se demuestra que cada vez se produzcan menos fallecidos en cada accidentes individual.

Como última puntualización, aunque en este análisis no afecta a las hipótesis, en otros análisis habría que tener muy en cuenta la evolución del número total de pasajeros transportados. Por ejemplo, en 2016 el Aeropuerto de Barajas recibió 50 millones de pasajeros, mientras que en 2004 fueron 38 millones

² Como conclusión añadida, me interesaría estar en el año más avanzado posible, pero no porque los accidentes sean menos dañinos, sino porque hay menos accidentes.

Recursos

- <https://relopezbriega.github.io/blog/2014/05/28/python-librerias-esenciales-para-el-analisis-de-datos/>
- <https://plot.ly/python/normality-test/>
- <https://dlegorreta.wordpress.com/2015/09/23/regresion-en-python/>
- http://www.ugr.es/~bioestad/_private/Tema_8.pdf
- <http://www.lsta.upmc.fr/sangnier/files/5MS101/4-Statistics.html>
- <http://nbviewer.jupyter.org/github/jvns/pandas-cookbook/blob/v0.1/cookbook/Chapter%201%20-%20Reading%20from%20a%20CSV.ipynb>
- <https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.bartlett.html>
- <https://blog.adrianistan.eu/2017/11/15/estadistica-python-analisis-datos-multidimensionales-regresion-lineal-parte-iv/>
- <https://support.minitab.com/es-mx/minitab/18/help-and-how-to/statistics/basic-statistics/how-to/correlation/interpret-the-results/key-results/>
- <http://www.aena.es/csee/Satellite?pagename=Estadisticas/Home>