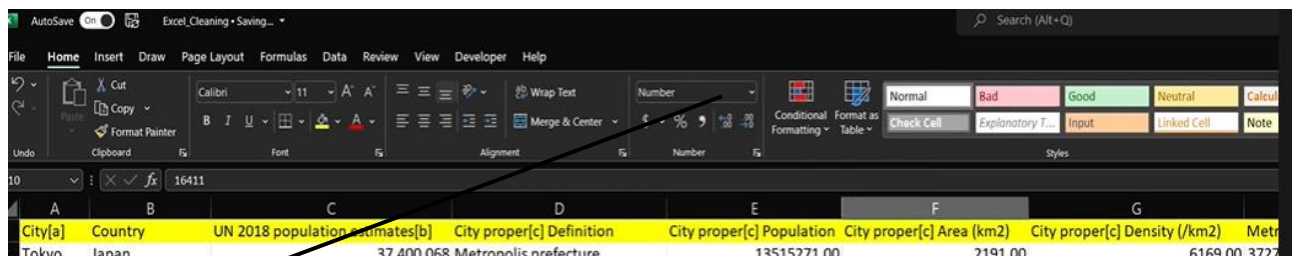


Types of values		
Categorical	Numerical	Composite
Boolean (True, False)	Integer (-2, -1, 0, 1, 2, ...)	Date / Time
Unordered (Red, Blue, Green)	Real (2.7, 3.1, ...)	Spatial (Lat/Long, Shapes)
Ordered (Low, Medium, High)		Structured (JSON, XML)
Unstructured (Text, Binary)		Specialized (IP, currency)

composite values are superset of numerical and categorical values.

- **Clean up data in excel:**
Find and replace a term: ctrl + h to open the pop up for the commands

Changing the data type:



The drop down will help you to change the data type of the column you have selected.

Remove extra spaces by using TRIM () function:

City proper[c] Definition	Trim_Colmn	City proper[c] Definition
Metropolis prefecture	=TRIM(D2)	Metropolis prefecture
Capital City		Capital City
Special city		Special city
Municipality		Municipality
Municipality		Municipality
City - state		City - state
Urban governorate		Urban governorate
Municipality		Municipality
Municipality		Municipality

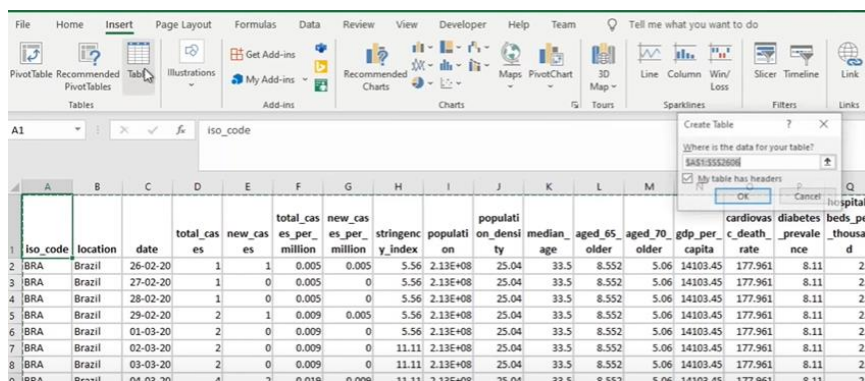
Selecting blank spaces (no entry in that row) and deleting rows: ctrl + G or select and replace drop down, it will then highlight the blank rows, then select the row and delete it

Remove duplicates: data tab. Function duplicate values

Split the columns: using text-to-columns in the data tab

Data aggregation: converting dataset to table-> insert tab -> table option

We convert dataset to a table because table has some very good features, it automatically comes with filter options. And when you try to enter a formula for any columns, you do not have to drag the formula for all the rows, it automatically does it for the entire column.



iso_code	location	date	total_cas	new_cas	total_cas_per_million	new_cas_per_million	stringency_index	population	population_density	median_age
BRA	Brazil	26-02-20	1	1	0.005	0.005	5.56	2.13E+08	25.04	33.5
BRA	Brazil	27-02-20	1	0	0.005	0	5.56	2.13E+08	25.04	33.5
BRA	Brazil	28-02-20	1	0	0.005	0	5.56	2.13E+08	25.04	33.5
BRA	Brazil	29-02-20	2	1	0.009	0.005	5.56	2.13E+08	25.04	33.5
BRA	Brazil	01-03-20	2	0	0.009	0	5.56	2.13E+08	25.04	33.5
BRA	Brazil	02-03-20	2	0	0.009	0	11.11	2.13E+08	25.04	33.5
BRA	Brazil	03-03-20	2	0	0.009	0	11.11	2.13E+08	25.04	33.5
BRA	Brazil	04-03-20	4	2	0.019	0.009	11.11	2.13E+08	25.04	33.5
BRA	Brazil	05-03-20	4	0	0.019	0	11.11	2.13E+08	25.04	33.5
BRA	Brazil	06-03-20	13	9	0.061	0.042	11.11	2.13E+08	25.04	33.5
BRA	Brazil	07-03-20	13	0	0.061	0	11.11	2.13E+08	25.04	33.5
BRA	Brazil	08-03-20	20	7	0.094	0.033	11.11	2.13E+08	25.04	33.5
BRA	Brazil	09-03-20	25	5	0.118	0.024	11.11	2.13E+08	25.04	33.5

Extract week from the date data by weeknum() function
month "mmm" and year "yyy" from the date data by using the formula text() formula

color scale: is a feature that help us identify clusters.

Select the column-> conditional formatting in the home tab->color scale-> more rules

2020	587646.4	13371	28776	166014
2020	5911758	35294	30393.29	166699
2020	5945849	34091	28312.71	167455
2020	5981767	35918	28597.86	168061
2020	602016.4	38397	29930.29	168613
2020	6052786	32622	29118.14	168989
2020	6071401	18615	29758.29	169183
2020	6087608	16207	30163.43	169485
2020	6118708	31100	29564.29	170115
2020	6166606	47898	31536.71	170769
2020	6204220	37614	31779	171460
2020	6238350	34130	31169.43	171974
2020	6290272	51922	33926.57	172561
2020	6314740	24468	34762.71	172833
2020	6335878	21138	35467.14	173120
2020	6386787	50909	38297	173817
2020	6426650	49863	38577.71	174515
2020	6487084	50434	40409.14	175270
2020	6533968	46884	42231.14	175964
2020	6577177	43209	40986.43	176628
2020	6603540	26363	41257.14	176941
2020	6623911	20371	41147.57	177317
2020	6674999	51088	41173.14	178159
2020	6728452	53453	41686	178995
2020	6781799	53347	42102.14	179765
2020	6836227	54428	43179.86	180437
2020	6880127	43900	43278.57	181123
2020	6901952	21825	42630.29	181402
2020	6927145	25193	43319.14	181835
2020	6970034	42889	42147.86	182799
2020	7040608	70574	44593.71	183735
2020	7110434	69826	46947.86	184827
2020	7162978	52544	46678.71	185650
2020	7213155	50177	47575.43	186356
2020	7238600	25445	48092.57	186764
2020	7263619	25019	48067.71	187291
2020	7318821	55202	49826.71	188259
2020	7365517	46696	46415.57	189220
2020	7423945	59428	44787.29	189982
2020	7448560	24615	40797.43	190488
2020	7465806	17246	36093	190795
2020	7484215	18479	35097.86	191139

for high values the colour is closer to red and for low it's yellow

Data bars: gives better representation of the values in the

Feb		
Mar	5517	1341
Apr	81470	33466
May	427662	155746
Jun	887192	394872
Jul	1260444	1110507
Aug	1245787	1995178
Sep	902663	2621418
Oct	724670	1871498
Nov	800273	1278727
Dec	1340095	803865
2021	10290858	20353258
Jan	1528758	490936
Feb	1346528	354631
Mar	2197488	1109424
Apr	1910264	6943304
May	1886543	9010075
Jun	1421277	2236590
Jul		208298
blank		

column

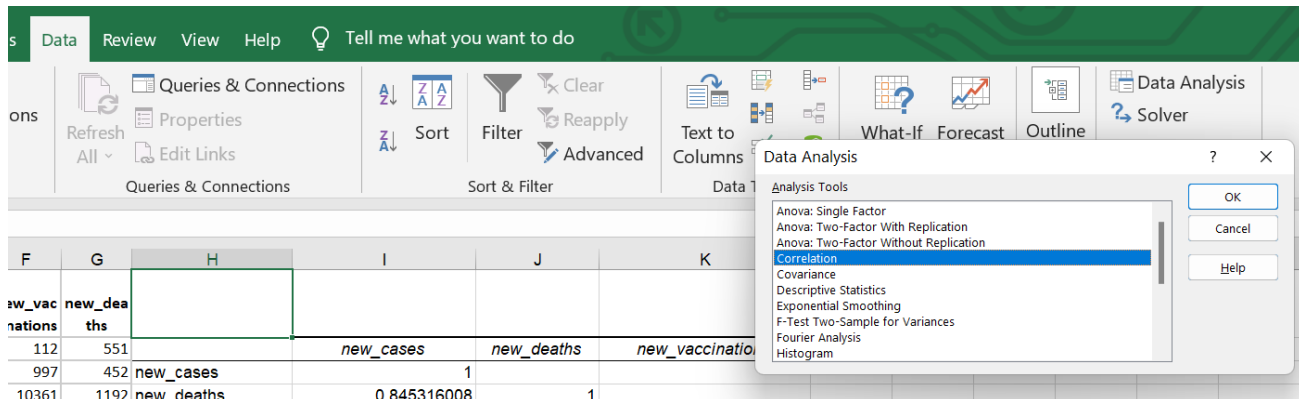
pandas_profiling library:

So, pandas profiling is one of the widely used libraries in the industry, which gives insights or the early findings when the data set which we are going to use for building any model.

from pandas_profiling import ProfileReport

OpenRefine: So, in this way, where we have different versions of the same entity, we can use the clustering algorithm of the Open Refine and clean the data and save them. For eg: XYZ_Ltd and XYZ_limited are the same companies for us but not for the computer. We use OpenRefine to select such data and change to same name.

Correlation using excel:



The screenshot shows the Microsoft Excel interface with the 'Data Analysis' dialog box open. The 'Correlation' option is selected under the 'Data Analysis' tab. The background shows a spreadsheet with columns for new_cases, new_deaths, and new_vaccinations, and a correlation matrix.

	new_cases	new_deaths	new_vaccinations
new_cases	1		
new_deaths	0.845316008	1	
new_vaccinations	0.288554403	0.235268388	1

Choose the data analysis option in data tab for the pop up. From there choose correlation option. Choose the columns for which to check the correlation.

	new_cases	new_deaths	new_vaccinations
new_cases	1		
new_deaths	0.845316008	1	
new_vaccinations	0.288554403	0.235268388	1

The correlation chart will appear then represent it using scatter plot.

Regression using Excel:

Linear regression is used to model the relationship between an independent variable and one or many dependent variables. If it is just one independent variable, then it is a simple linear regression, but if it involves many independent variables, it is multiple linear regression. Use the similar correlation processes.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.904484662							
R Square	0.818092504							
Adjusted R Square	0.816493317							
Standard Error	479.4514205							
Observations	460							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	4	470383883.9	117595971	511.5678263	7.729E-167			
Residual	455	104592517.4	229873.6647					
Total	459	574976401.3						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-235.9546149	98.78576468	-2.388548752	0.017321792	-430.0875529	-41.82167696	-430.0875529	-41.82167696
new_cases_per_thousand	7.248436244	0.360266161	20.11966994	7.19878E-65	6.540444273	7.956428215	6.540444273	7.956428215
new_tests_per_thousand	0.691822498	0.05952635	11.62212192	1.60744E-27	0.574841825	0.808803171	0.574841825	0.808803171
new_vaccinations_per_thou	-0.070876801	0.024544361	-2.88770201	0.004065628	-0.119111168	-0.022642433	-0.119111168	-0.022642433
stringency_index	2.713139528	1.722286555	1.575312494	0.115879348	-0.671483246	6.097762302	-0.671483246	6.097762302

R square:

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination.

value ranges between 0-1. The value tells how much of the data (dependent variable) can the model (built up by independent vars) explain. Here, r square is 0.8180 which means the model can explain 0.81 or 81 % of the data, which also means that the mode cannot explain 19 % of the data

The more variance that is accounted for by the regression model the closer the data points will fall to the fitted regression line. Theoretically, if a model could explain 100% of the variance, the fitted values would always equal the observed values and, therefore, all the data points would fall on the fitted regression line

p-value:

HOW DO I INTERPRET THE P-VALUES IN LINEAR REGRESSION ANALYSIS?

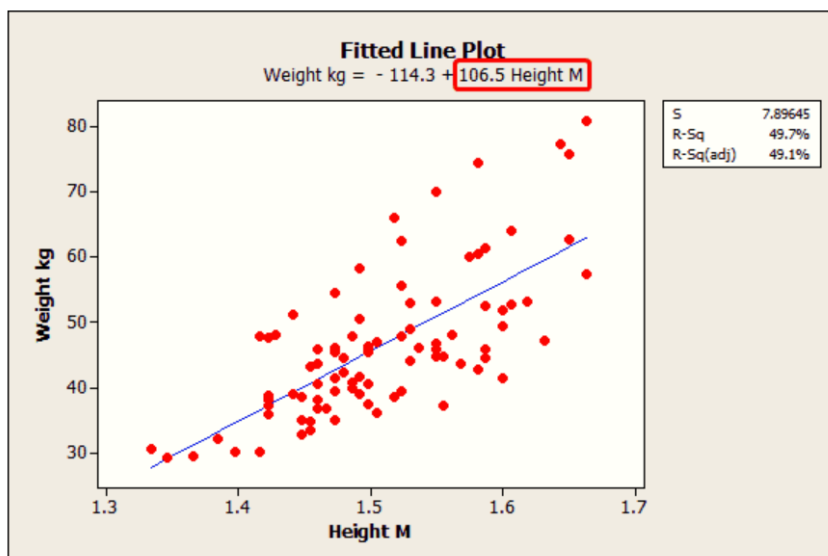
The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, an independent var that has a low p-value is likely to be a meaningful addition to your model because changes in the independent var's value are related to changes in the dependent variable.

Conversely, a larger (insignificant) p-value suggests that changes in the independent var are not associated with changes in the dependent variable.

In the above example we have all the independent var's p-value < 0.05 . Therefore, we see that all the predictors are related to the response var.

Regression coefficients:

Regression coefficients represent the mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant.



Coefficients

Term	Coef	SE Coef	T	P
Constant	-114.326	17.4425	-6.55444	0.000
Height M	106.505	11.5500	9.22117	0.000

The blue fitted line graphically shows the same information. If you move left or right along the x-axis by an amount that represents a one-meter change in height, the fitted line rises or falls by 106.5 kilograms $\Rightarrow y = mx + c \Rightarrow \text{Weight} = 106.505 \text{ height} + c$

If the fitted line was flat (a slope coefficient of zero), the expected value for weight would not change no matter how far up and down the line you go. So, a low p-value suggests that the slope is not zero, which in turn suggests that changes in the predictor variable are associated with changes in the response variable

P-value:

Imagine the following:

- You have an hypothesis that you want to disprove (reject)
 - That is normally called the null hypothesis (or H_0)
- To try to do that, you run a statistical test (like a t-test) on your data
- That test will return a value, the t statistic
 - Ignore what the t-statistic is – it would take longer to explain, just consider it is the answer to your initial question;
- You have to determine “how confident” you want to be when rejecting the null hypothesis
 - That is your level of significance (also called alpha)
 - “normally” it is 95%, or 0.05
- The level of significance will determine another value, called the “t-critical” value determined from the t-table.
 - For the sake of simplicity, let’s assume a two-tail test – you are testing if your value is “different” (can be smaller or bigger) – so divide your level of significance by two – 2.5% instead of 5% on each tail
- If your t-statistic is bigger than the t-critical or smaller than negative t-critical, you have enough evidence to reject the null hypothesis;
- The p-value, will be the area under the curve from your negative t statistic to minus infinity or your positive t-statistic plus infinity, depending on which side of the t-distribution you are;
 - That is why, the smaller the p-value the bigger the evidence against H_0

Example:

You have a process that normally produces something (choose whatever you want) that weights 5g

One day you decide to sample 10 items from that process and take the average of their weights

The average is 6g with a standard deviation of 1g, so you wonder: is something “wrong” (or different) with my process?

Your null hypothesis is: no, there is nothing wrong

Your alternative hypothesis is: yes, there is something wrong with my process

So you choose your confidence level (95% – 2.5% on each tail)

You run your test and get t statistic: 3.162

Given your alpha and your number of samples, the t critical is ± 2.26 found from the t-table

Since your t-statistic is bigger than your t-critical, you reject the null hypothesis

Your p-value will be the area under the t-curve from $t = 3.162$ to infinity, in this case: 0.01150 (which, of course, is smaller than 0.025) therefore $p\text{-value} < 0.025(\alpha)$ gives evidence for rejection of null hypothesis.

ANOVA and F-test:

The ANOVA test allows a comparison of more than two groups at the same time to determine whether a relationship exists between them. The result of the ANOVA formula, the F statistic (also called the F-ratio), allows for the analysis of multiple groups of data to determine the variability between the data and within the data.

$$F = \frac{\frac{\text{Sum of Squares Between Groups}}{\text{degrees of freedom numerator}}}{\frac{\text{Sum of Squares Within Groups}}{\text{degrees of freedom denominator}}}$$

If no real difference exists between the tested groups, which is called the [null hypothesis](#), the result of the ANOVA's F-ratio statistic will be close to 1.

$$\frac{\text{Sum of Squares Between Groups}}{\text{degrees of freedom}} = \frac{203.3}{2} = 101.667$$

$$\frac{\text{Sum of Squares Within Groups}}{\text{degrees of freedom}} = \frac{54}{12} = 4.5$$

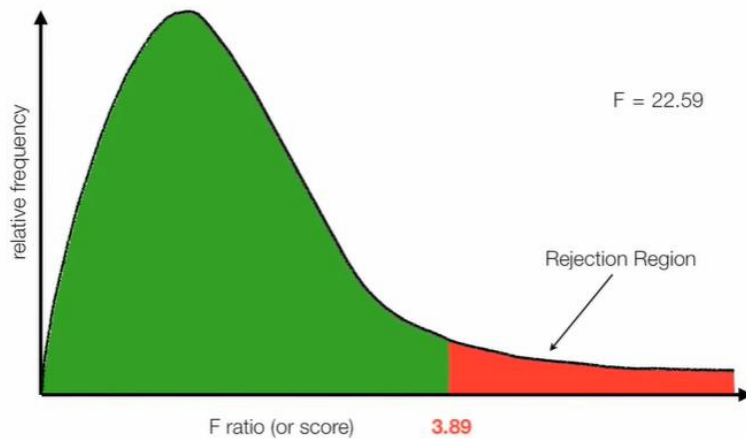
$$F = \frac{101.667}{4.5} = 22.59$$

Eg:

We then calculate the critical value which is found by using $F(\text{degrees of freedom numerator}, \text{degrees of freedom denominator})$ from the F table

		degrees of freedom numerator														
degrees of freedom denominator		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	1	161.5	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	246.0	248.0	249.1	250.1
	2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46
	3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62
	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75
	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50
	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81
	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38
	8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08
	9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86
	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70
	11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57
	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47
	13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38
	14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31
	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25

We then use the critical region to separate the acceptance and rejection area. We then check if our F-statistic value falls in the acceptance or rejection area



Here we reject the null- hypothesis \Rightarrow calculated value $>$ critical value. Also, since $p\text{-value} < \alpha (0.05) = \text{level of significance}$, there is a relationship between the groups.

F value and it's P value for regression

The F-statistic provides us with a way for globally testing if ANY of the independent variables $X_1, X_2, X_3, X_4, \dots$ is related to the outcome Y.

For a significance level of 0.05:

- If the p-value associated with the F-statistic is ≥ 0.05 : Then there is no relationship between ANY of the independent variables and Y
- If the p-value associated with the F-statistic < 0.05 : Then, AT LEAST 1 independent variable is related to Y

F-test P-value and individual independent var's p-value, which one to check in the regression output table?

Why do we even need the F-test?

A T-test will tell you if a *single* variable is statistically significant and an F test will tell you if a *group* of variables are jointly significant.

Why do we need a global test? Why not look at the p-values associated with each coefficient $\beta_1, \beta_2, \beta_3, \beta_4, \dots$ to determine if any of the predictors is related to Y?

Notice that the coefficient of X_3 has a p-value < 0.05 which means that X_3 is a statistically significant predictor of Y:

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4)

Residuals:
    Min       1Q   Median       3Q      Max
-2.30084 -0.63789 -0.00759  0.62416  2.94896

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.13027    0.09503   1.371  0.1736
x1           0.05365    0.09730   0.551  0.5826
x2          -0.01173    0.10052  -0.117  0.9073
x3           0.19944    0.09756   2.044  0.0437 *
x4           0.07307    0.09840   0.743  0.4596
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9147 on 95 degrees of freedom
Multiple R-squared:  0.05495,    Adjusted R-squared:  0.01516
F-statistic: 1.381 on 4 and 95 DF,  p-value: 0.2464
```

p-values



However, the last line shows that the F-statistic is 1.381 and has a p-value of 0.2464 (> 0.05) which suggests that NONE of the independent variables in the model is significantly related to Y! So is there something wrong with our model? If not, then which p-value should we trust: that of the coefficient of X_3 or that of the F-statistic?

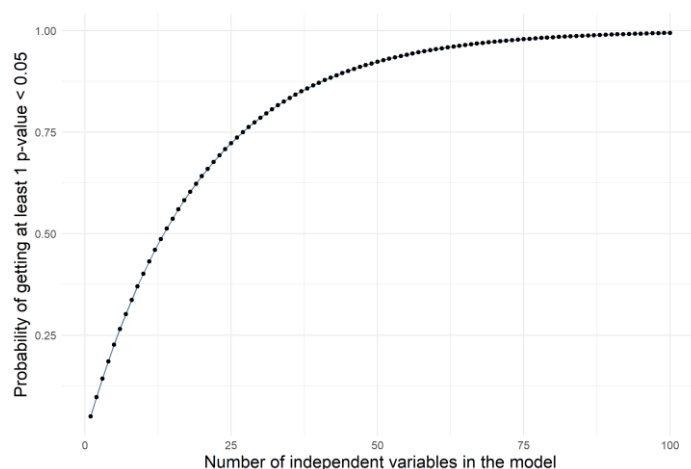
The answer is that we cannot decide on the global significance of the linear regression model based on the p-values of the β coefficients.

This is because each coefficient's p-value comes from a separate statistical test that has a 5% chance of being a false positive result (assuming a significance level of 0.05).

For instance, if we take the example above, we have 4 independent variables (X_1 through X_4) and each of them has a 5% risk of yielding a p-value < 0.05 just by chance (when in reality they're not related to Y).

The more variables we have in our model, the more likely it will be to have a p-value < 0.05 just by chance.

Here's a plot that shows the probability of having AT LEAST 1 variable with p-value < 0.05 when in reality none has a true effect on Y:



In the plot we see that a model with 4 independent variables has a 18.5% chance of having at least 1 β with p-value < 0.05 .

The plot also shows that a model with more than 80 variables will almost certainly have 1 p-value < 0.05 .

Therefore it is obvious that we need another way to determine if our linear regression model is useful or not (i.e. if at least one of the

X_i variables was important in predicting Y). Here's where the F-statistic comes into play.

One important characteristic of the F-statistic is that it adjusts for the number of independent variables in the model. So it will not be biased when we have more than 1 variable in the model.

Returning to our example above, the p-value associated with the F-statistic is ≥ 0.05 , which provides evidence that the model containing X_1, X_2, X_3, X_4 is not more useful than a model containing only the intercept β_0 .

BOTTOM LINE:

In this example, according to the F-statistic, none of the independent variables were useful in predicting the outcome Y , even though the p-value for X_3 was < 0.05 .

What if the F-statistic has a statistically significant p-value but none of the coefficients does?

Here's the output of another example of a linear regression model where none of the independent variables is statistically significant but the overall model is (i.e., at least one of the variables is related to the outcome Y) according to the p-value associated with the F-statistic.

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4)

Residuals:
    Min       1Q   Median       3Q      Max
-2.4847 -0.5768  0.1114  0.7021  2.3773

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.08179   0.10763   -0.760   0.4492
x1           0.24876   0.14063    1.769   0.0801 .
x2           0.19768   0.11076    1.785   0.0775 .
x3           0.13203   0.10542    1.252   0.2135
x4           0.03620   0.10530    0.344   0.7318
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.066 on 95 degrees of freedom
Multiple R-squared:  0.1317,    Adjusted R-squared:  0.09516
F-statistic: 3.603 on 4 and 95 DF,  p-value: 0.008853
```

But how is that even possible?

Well, in this particular example I deliberately chose to include in the model 2 correlated variables: X_1 and X_2 (with correlation coefficient of 0.5).

Because this correlation is present, the effect of each of them was diluted and

therefore their p-values were ≥ 0.05 , when in reality they both are related to the outcome Y .

CONCLUSION:

When it comes to the overall significance of the linear regression model, always trust the statistical significance of the F-statistic over that of each independent variable.

Outliers: calculate Q1, Q3 using QUARTILE.EXC() function, calculate IQR = Q3 - Q1

Lower bound = Q1 - 1.5(IQR), upper bound = Q3 + 1.5(IQR)

If $\text{data_point} < \text{lower bound}$ or $\text{data_point} > \text{upper bound}$ then it is an outlier

ARIMA: