

Question 1

For the variable month: We select the column Month and then:

Insert Delete Format ∑ AutoSum ∑
Image: Delete Format ∑ AutoSum ∑ Point Insert Delete Format Fill * Sort & Find & Fill * Cells Clear * Ø Find Ø O To Go To Go To O To Go To Go To O To Cognents Cognents Cogstants Dot Voltage Dat Voltage Data Voltage Dot Voltage Dot Voltage
Cells Ed Ø Eind Ø Go To Ø Go To Ø O Formylas Cogments Cogments Constants Dats yelidab Dats Able Objectab Dats Able Objectab
N O So To Go To Go To Special. Formylas Comments Comments Conditional Fe Data Validation Data Validation Data Validation
→ Go To Go To Special. Formylas Cogments
N O Formulas Cogments Conditional Formulas Codments Conditional Formulas Codditional Formulas Data Validation Data Validation
N O Comments Conditional Fo Cogstants Data Validation
Constituent Fice
Constants Data Validation Data Validation
Data Validation
Calact Objects
Mr. Select Objects
Et Selection Pane
and Replace Fing Replace
Replace with: May
Replace All Replace Find All Find Nex

(we do the replacement for each month).

For the variable TempCelsius: For the transformation from Fahrenheit (F) degrees to Celsius (C) degrees we use the formula:

$$C = (F - 32) \cdot \frac{5}{9}$$

So, in Excel for 67 degrees Fahrenheit we have 19.44 degrees Celsius from the formula: (D2-32)*5/9

For the variable HotTemp: We use the formula in Excel: IF(G2>25;1;0)

For the variable OzAlertLevel: We use the formula in Excel:

IF(AND(A2>65;G2>25);2;IF(AND(A2>50;A2<=65;G2>25);1;0))

Question 2

2.1 We use the function AVERAGE for the mean and the function STDEV.S for the standard deviation (see excel file).



	T	empCelsius
	Mean	Standard Deviation
Мау	18.6	3.81
June	26.2	3.67
July	28.8	2.40
August	28.9	3.66
September	24.9	4.64

The largest temperature in degrees Celsius is noted in August. The largest standard deviation is noted in August.

2.2 We create the variable **"Hot days"** in Excel by using the IF function (IF(Cell>25;"Hot day";"Not hot day")). Then we use the function COUNTIF to count the hot days. To calculate the percentage of hot days for each month, we divide the number of hot days with each month's total days.

	May	June	July	August	September
Number of Hot days	2	16	29	26	12
Percentage	0.06	0.53	0.94	0.84	0.4 🥥

The largest percentages of hot days are noticed in June (53%), July (94%) and August (84%). These findings are in line with the ones from question 2.1, as the largest means are also noticed in June, July and August.

2.3 From Excel, by using MEDIAN and QUARTILE functions:

Variable Wind	May	June	July	August	September
Median	11.5	9.7	8.6	8.6	10.3
1st quartile	8.9	8	6.9	6.6	7.55
3rd quartile	14.05	11.5	10.9	11.2	12.325

The mean of Wind in August is 8.79 which is larger than the median, indicating a right skewed distribution.





2.4 For comparing the variability of different sets of data, we will use the Coefficient of variation (CV). CV is defined as:

$$CV = \frac{s}{\bar{x}}$$

Variable	Coefficient of variation
Ozone	0.73
TempCelsius	0.21
Solar.R	0.47
Wind	0.35

Ozone variable shows the largest variability, as it has the largest coefficient of variation (73 %).

2.5 The number of classes k is calculated from Sturge's rule:

$$k = 1 + 3.322 \cdot \log(n) = 1 + 3.322 \cdot \log(153) = 8.2575 \approx 8$$

To calculate the frequencies of the frequency table we use the COUNTIFS function in Excel. For example for the frequency of the first class [0,42], we have COUNTIFS($F^2:F^{154};">=0";F^2:F^{154};"<=42"$)

Solar.R classes	Upper Class Bound	Class midpoint (m _i)	Frequency (fi)	Frequency (%)	mi*fi
[0,42]	42	21	13	8%	273
(42,84]	84	63	15	10%	945
(84,126]	126	105	11	7%	1.155
(126,168]	168	147	16	10%	2.352
(168,210]	210	189	26	17%	4.914
(210,252]	252	231	29	19%	6.699
(252,294]	294	273	32	21%	8.736
(294,336]	336	315	11	7%	3.465
Total			153	100%	28539

The mean in the grouped data is:

$$\bar{x} = \frac{\sum m_i f_i}{n} = \frac{28539}{153} = 186.53$$

The **mean in the raw data** is:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{28386}{153} = 185.53$$



0

1

If the raw data is available, then we better use the raw data mean, a measure that utilizes all the available information from the data.

Question 3

3.1 The number of classes k is calculated from Sturge's rule:

 $k = 1 + 3.322 \cdot \log(n) = 1 + 3.322 \cdot \log(153) = 8,2575 \approx 8$

The frequency table is:

Ozone classes	Upper Class Bound	Frequency	Frequency
[0,21]	21	41	27%
(21,42]	42	58	38%
(42,63]	63	25	16%
(63,84]	84	15	10%
(84,105]	105	7	5%
(105,126]	126	5	3%
(126,147]	147	1	1%
(147,168]	168	1	1%
Total		153	100%

For the **frequency histogram** we select the classes column and the frequency column. Then:

				αρχειο excel - Excel
		Store	2 🚮 📭	1 ▼ ■ ▼ ☆ ▼ 1 ■ 1
PivotTable Recommended Table PivotTables Tables	Pictures online snapes smarcArt sci Pictures * Illustrations	Add-ins	aps Graph Charts	
Chart 8 🔻 🗄 🗡				3-D Column
E	FG	H	1	
1				41
				More Column Charts

Then we double click on the bars and we set Gap Width to 0 %.







We observe that the distribution of Ozone is **skewed to the right** (positive skewness) and there exists one peak, which is the mode (29 if we calculate it with the MODE function in Excel).

In Excel (version 2016 and above), select the Insert option:

E	5 •∂~∶	≂							mba602021wa1dat	a - Excel
Fil	le Home	Insert Pa	ige Layout	Formulas	Data Review	View	Help Pow	er Pivot G	Tell me what you want to	o do
Pivo	otTable Recomme PivotTab	ended Table	Pictures	🖥 Online Pictur 🖓 Shapes 🕶 🖹 SmartArt	res 👩 🔒 Screenshot 👻	🕂 Get /	Add-ins V Add-ins V	Recomme Chart	nded s Histogram	atChart 2D
	Tables			Illustrati	ons	1	Add-ins		Cha the Pr	<u> </u>
B1	· · ·	× ✓	f _x A	/C_AGE						
	А	В	С	D	E		F	G	H Box and Wei	ker
1	A/C_YEAR	A/C_AGE								
2	2018	2			Minimum		0			
3	2018	2			Maximum		28		min <u>In</u> ore Stat	istical Charts
bo	xplot is	:								6
				E	Boxplot	of O	zone			\mathbf{G}
-	180				-					and the second s
	160					•				

We notice the presence of 4 outliers. The distribution is skewed to the right. 3.2 In Excel (version 2016 and above) we choose Insert then Box and Whisker:

E	ქ რ ო (⇒ - ≂			mba	602021wa	1data - Exce	2l			Char	t Tools		
F	ile Hor	me Inse	ert Page	Layout	Formulas	Data	Review	View	Help Po	ower Pivot	Design	Format	Q Tell me what yo	ou wa
[Piv	otTable Reco	ommended votTables	Table	Pictures	Online Pictu Shapes + SmartArt	res 👩 S	creenshot *	Get /	Add-ins Add-ins 🔻	Reco	mmended Charts		gram	
Ch	T. art 1	ables	X V	f_{x}	Illustrat	ions			Add-ins			Cha		
	Α	в	с	D	E	F	G	н	I	J	к	Box	and Whisker	
1	4000	7000	8000	15000	15000								i.	ĺ
2	1000	5000	5000	6200	5000							ΓΫ́Ω	<u>¥</u>	
3	4000	2000	3000	15000	10000		0					- Idla M	Anne Statistical Charts	-
4	2000	4300	10000	8000	5000		Ύ.	0000				<u></u>	nore statistical charts	- L

The boxplots for each month are:





By examining the box plots, we can identify that:

- The distributions for Ozone in July and August are positively skewed.
- The distributions for Ozone in May and September are not skewed.
- The IQR value is larger in July and August.
- Outliers exist in May, June, August and September.
- The median Ozone in May and September is lower than the grand median (the grand median is 29).

3.3 To construct the scatter plots we select each pair of 2 variables and then:



The scatterplots are:







<u>Negative linear relationship</u> exists in the pairs of variables (wind, temp), (ozone, wind).

<u>Positive linear relationship</u> exists in the pair of variables (temp, ozone).

3.4 We construct the following table in Excel, by using the COUNTIF function:

OzAlertLevel	0	1	2
May	30	0	1
June	29	0	1
July	11	11	9
August	14	6	11
September	26	0	4

For the clustered bar chart, we select the table in Excel and then:



File	Home	Insert Page L	ayout Formulas	Data Re	view View	Help F	Power Pive	ot Q T	ell me what	you want t	to do	
Pivot	able Recommen	ded Table P	Online Pi	ctures 👩 🛊 Screer	nshot - 🗄 G	Get Add-ins	🄹 🛛		2-D Colu	• 1 % •		0.10
	PivotTable Tables	es	SmartArt Illus	rations	• •	Add-ins		Charts				
H98	• E	$\times \checkmark$	fx						3-D Colu	imn		
									40	þ₿	J.	JJ.
	А	В	С	D	E			F	2-D Bar			
96	16	August	82	7,4		0)					
97	78	August	86	6,9	0	2	2					
98	35	August	85	7,4	Ĭ				3-D Bar			2
99	66	August	87	4,6						B	B	
100	122	August	89	4	35 —					e Column (Charts	
					- I					ecolumn	enarcs	

The clustered bar chart is:



Ozone > 65 and TempCelsius > 25 are mainly noticed in July and August. Ozone > 50 & Ozone \leq 65 & TempCelsius > 25 are also mainly noticed in July and August.

Question 4

We first construct the variable "Yellow flag" by using the function IF(cell of Ozone>65;1;0)

The probability of a yellow flag in any given day is p=0.17 or 17 % (see excel file).

Number of yellow flags	26
Percentage of yellow flags	0.17

Let *X* the random variable that denotes the number of yellow flags in a week. Then *X* follows a Binomial distribution with n = 7 and p = 0.17. Hence:



$$X \sim B(n = 7, p = 0, 17)$$

The probability that at least one yellow flag within a week (7 days) is between May and September is:

$$P(X \ge 1) = 1 - P(X < 1) = 1 - P(X = 0) = 1 - 0.27 = 0.73$$

We calculate the probability P(X = 0) with the help of Excel.

Input value	0
Number of trials	7
Event probability	0.17
P(X=0)	0.27
1-P(X=0)	0.73

4.2 The expected number of yellow flags during the same week is:

$$E(X) = n \cdot p = 7 \cdot 0.17 = 1.19 \approx 1$$

Question 5

5.1 We first construct the variable "Red flag" by using the function IF(AND(cell of ozone>65;cell of temp>90);1;0).

The probability of a red flag in any given day is p=0.065 or 6.5 % (see excel file).

Number of red flags	10
Percentage of red flags	0.065

We denote as X the random variable number of red flags within 31 days. For large values of n and small values of p, the Poisson distribution approximates the binomial distribution. The assumptions are:

$$n > 20, np < 5 \text{ or } n(1-p) < 5$$

Hence:

$$n = 31 > 20$$
, $np = 31 \cdot 0.065 = 2.015 < 5$

As a result, the Binomial distribution can be approximated by the Poisson distribution. The parameter of the Poisson distribution is:

$$\lambda = n \cdot p = 2.015$$

The probability of at least two red flags within a given month is:

 $P(X \ge 2) = 1 - P(X < 2) = 1 - P(X \le 1) = 1 - 0.4020 = 0.5980$



Poisson Mean	2.015
Probability P(X<=1)	0.4020
1-P(X<=1)	0.5980

Note: We can calculate $P(X \le 1)$ *by using the* POISSON.DIST(1;2,015;TRUE) *function in Excel.*

5.2 The probability of at least one but no more than three red flags during the same month is:

$$P(1 \le X \le 3) = P(X = 1) + P(X = 2) + P(X = 3) = 0.269 + 0.271 + 0.182$$

= 0.722

P(X=1)	0.269
P(X=2)	0.271
P(X=3)	0.182

The function we used for the above probabilities is:

POISSON.DIST(x;2,015;FALSE) where x = 1,2,3.

<u>Question 6</u>

6.1 Let X denotes the temperature. According to the exercise and Excel, X follows normal distribution with parameters $\mu = 25.5$ and $\sigma = 5.26$.

We want to calculate the probability

P(X > 20)

We calculate the probability $P(X \le 20)$ with the help of Excel:

Mean (µ)	25.5
Standard deviation (σ)	5.26
Probability P(X<=20)	0.148

The function we used for the previous cumulative probability is: NORM.DIST(20;D3;D4;TRUE). So:

$$P(X > 20) = 1 - P(X \le 20) = 1 - 0.148 = 0.852$$

6.2 Let us denote the minimum temperature that a day must have (at least) in order to be among the 10% of the warmest days of the season, as x_0 . Then the requested probability is:

$$P(X > x_0) = 0.10 \Longrightarrow 1 - P(X \le x_0) = 0.10 \Longrightarrow P(X \le x_0) = 0.90$$

5

C_

in



Mean	25.5
Standard deviation	5.26
Inverse Probability	32.22921

To find x_0 we use NORM.INV(0,9;H3;H4) in Excel. This means that $x_0 = 32.22921$.