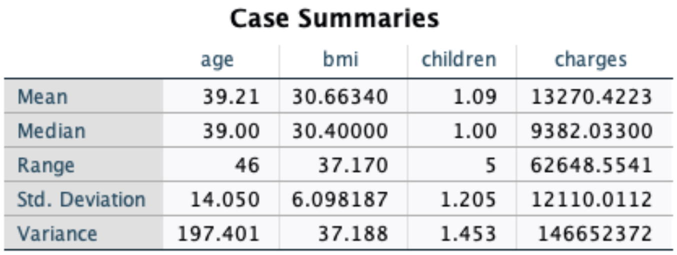
Solution_Insurance Charges Data Analysis

I. State the measurement level (nominal, ordinal, interval, ratio) for each variable in the data set along with the appropriate measures for central tendency (mode, mean, median) and dispersion (range, variance/standard deviation)

age: ratio. For a person, the age could be 0, which is meaningful.
sex: nominal. A categorical variable, there is no order between Male and female.
bmi: interval. It is a metric variable.
children: ratio. For an insurance customer, he or she could have no children.
smoker: nominal. Yes or no.
region: nominal. Northeast, northwest, southeast, and southwest.
charges: interval. It is a metric variable. But we cannot calculate the ratio between two different charges.

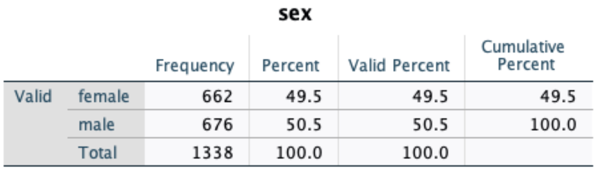
Metric:



Case Summaries

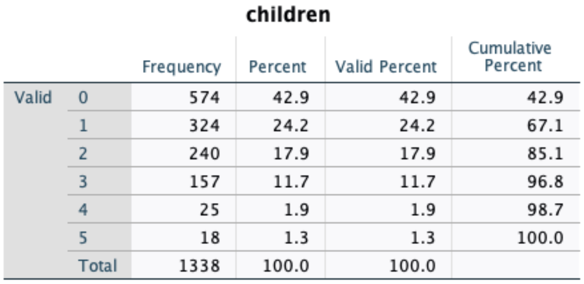
	age	bmi	children	charges
Mean	39.21	30.66340	1.09	13270.4223
Median	39.00	30.40000	1.00	9382.03300
Range	46	37.170	5	62648.5541
Std. Deviation	14.050	6.098187	1.205	12110.0112
Variance	197.401	37.188	1.453	146652372

Non-metric:



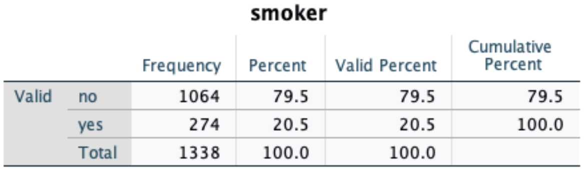
sex

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	female	662	49.5	49.5	49.5
	male	676	50.5	50.5	100.0
	Total	1338	100.0	100.0	



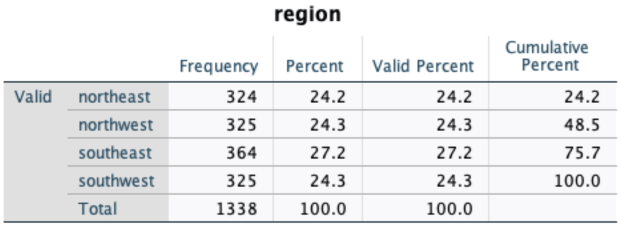
children

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	574	42.9	42.9	42.9
	1	324	24.2	24.2	67.1
	2	240	17.9	17.9	85.1
	3	157	11.7	11.7	96.8
	4	25	1.9	1.9	98.7
	5	18	1.3	1.3	100.0
	Total	1338	100.0	100.0	



smoker

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	no	1064	79.5	79.5	79.5
	yes	274	20.5	20.5	100.0
	Total	1338	100.0	100.0	



region

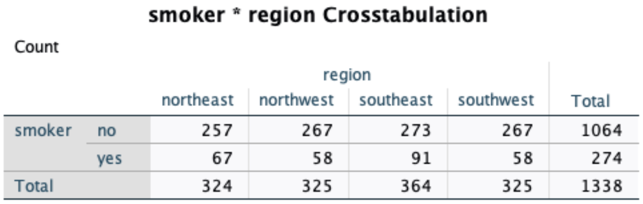
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	northeast	324	24.2	24.2	24.2
	northwest	325	24.3	24.3	48.5
	southeast	364	27.2	27.2	75.7
	southwest	325	24.3	24.3	100.0
	Total	1338	100.0	100.0	

II. For each of the questions a-c below select the appropriate test and state:

- What is the null hypothesis?
- Did you reject the null? Why or why not?
- What is your conclusion?

a. Is there a relationship between being a smoker and the region a person lives in?

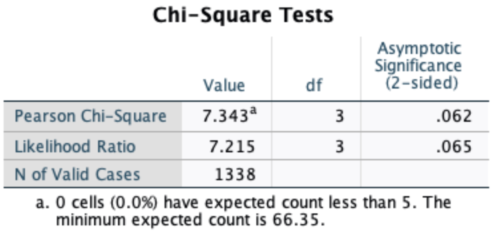
Chi-square test



smoker * region Crosstabulation

Count

		region				
		northeast	northwest	southeast	southwest	Total
smoker	no	257	267	273	267	1064
	yes	67	58	91	58	274
Total		324	325	364	325	1338



Chi-Square Tests

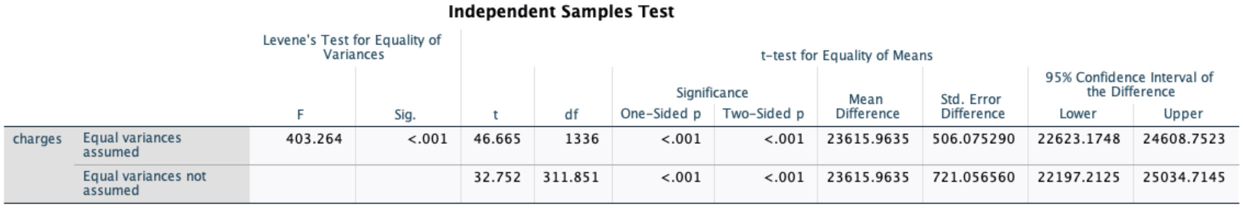
	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	7.343[a]	3	.062
Likelihood Ratio	7.215	3	.065
N of Valid Cases	1338		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 66.35.

H_0: There is no relationship between the region a person lives and being a smoker.
The P-value is 0.062 > 0.05, which means we cannot reject the null hypnosis.
Conclusion: whether a person smoking or not does not have association with the region a person lives in.

b. Are smokers charged more by insurers relative to non-smokers?

T-test



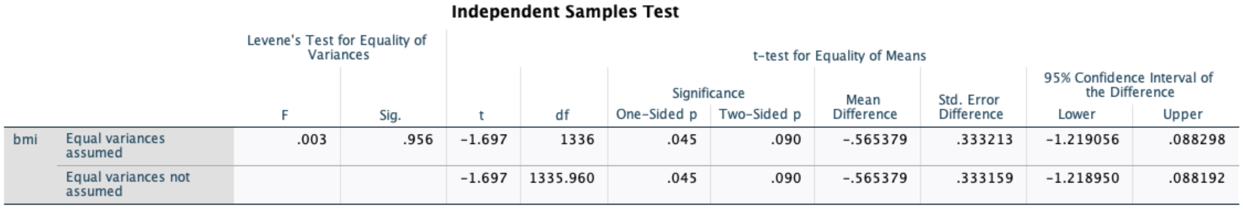
Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
						Significance		Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p			Lower	Upper
charges	Equal variances assumed	403.264	<.001	46.665	1336	<.001	<.001	23615.9635	506.075290	22623.1748	24608.7523
	Equal variances not assumed			32.752	311.851	<.001	<.001	23615.9635	721.056560	22197.2125	25034.7145

H_0: Smokers are not charged more than non-smokers.
As P-value < 0.05, we can reject H_0.
Conclusion: Insurers charge differently between smokers and non-smokers.

c. Is BMI different for males and females?

T-test



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
						Significance		Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p			Lower	Upper
bmi	Equal variances assumed	.003	.956	-1.697	1336	.045	.090	-.565379	.333213	-1.219056	.088298
	Equal variances not assumed			-1.697	1335.960	.045	.090	-.565379	.333159	-1.218950	.088192

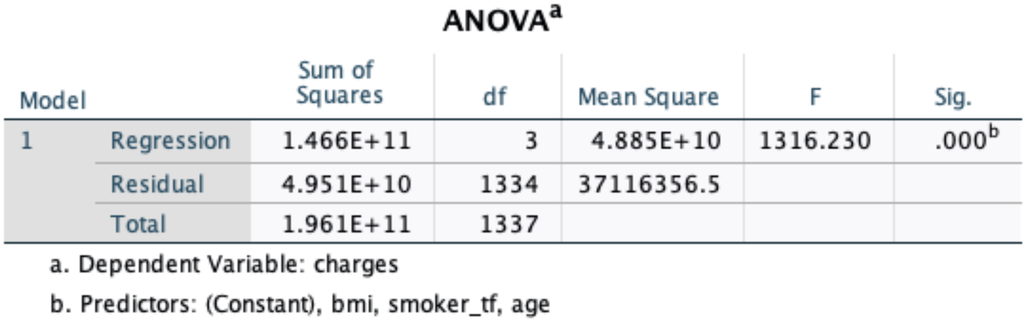
H_0 : BMI is not different for males and females.
We cannot reject H_0, because P-value > 0.05.
Conclusion: The BMI for males and females is similar.

III. Use a linear regression model to capture the relationship between insurance charges and relevant explanatory variables.

a. Briefly explain the rationale behind your model specification. Why did you include the variables you selected? Think about whether you need/want interaction effects or nonlinear transformation of the variables.

We started by including variables: bmi, age and smoker status in our model. In Q2 part, we know that smokers might be charged more than non-smokers. So, smoker factor needs to be included. Besides, variables of bmi and age have a significant effect on the overall health of the person which we would then assume they would have a strong correlation with the amount of charge.
For a linear regression, we do not want interaction effects and nonlinear transformation among variables. As any interact factors may influence the relationships among them. For instance, height or weight variables would influence bmi variable.

b. Comment on the overall fit of the model.



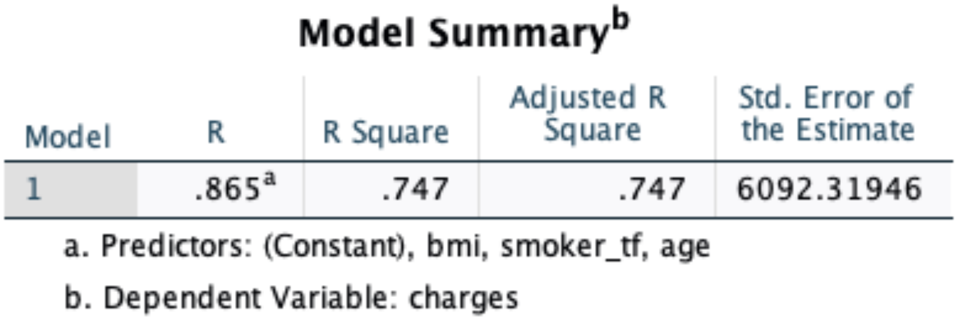
ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.466E+11	3	4.885E+10	1316.230	.000[b]
	Residual	4.951E+10	1334	37116356.5		
	Total	1.961E+11	1337			

a. Dependent Variable: charges

b. Predictors: (Constant), bmi, smoker_tf, age

This model is appropriate because the F equals 1316.23 which is far way larger than 3.



Model Summary[b]

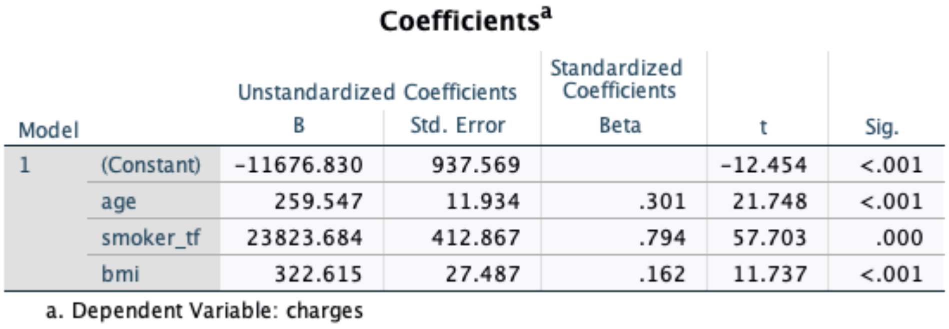
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.865[a]	.747	.747	6092.31946

a. Predictors: (Constant), bmi, smoker_tf, age

b. Dependent Variable: charges

In this model, there are 74.7% of cases could be explained.

c. Interpret the coefficients. What do we learn about the factors explaining the variation in insurance charges?



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-11676.830	937.569		-12.454	<.001
	age	259.547	11.934	.301	21.748	<.001
	smoker_tf	23823.684	412.867	.794	57.703	.000
	bmi	322.615	27.487	.162	11.737	<.001

a. Dependent Variable: charges

Coefficients reflect the linear relationships between charges, the dependent variable, and other independent variables.
age: when the customer's age increases one year, the insurance fee will increase by $259.547.
bmi: when the bmi increases one unit, the customer will cost $322.615 more.
smoker: The insurer would charge $23823.684 more fee to smokers than to non-smokers.

Overall, according to the degree of insurance fee variation, we can see that smoking factor is the main one influencing the fee.