Outcome of your EDA

Did multiple times with different variables from the data set and this is the best significance I found from the dataset.

OLS Regression Results

Dep. Variable:	DEATH_EVENT	R-squared:	0.072
Model:	OLS	Adj. R-squared:	0.069
Method:	Least Squares	F-statistic:	23.09
Date:	Sun, 06 Jun 2021	Prob (F-statistic):	2.45e-06
Time:	08:35:29	Log-Likelihood:	-185.33
No. Observations:	299	AIC:	374.7
Df Residuals:	297	BIC:	382.1
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.7253	0.088	8.235	0.000	0.552	0.899
EJECTION_FRACTION	-0.0106	0.002	-4.806	0.000	-0.015	-0.006

Omnibus:	147.000	Durbin-Watson:	1.114
Prob(Omnibus):	0.000	Jarque-Bera (JB):	43.084
Skew:	0.728	Prob(JB):	4.41e-10
Kurtosis:	1.843	Cond. No.	135.

Looking at the results we can say that null hypothesis can be rejected because the pvalue is less than 0.05. Therefore the variable has some impact on the death of the person due to heart failure.

What do you feel was missed during the analysis?

I felt like the data that was used to perform the analysis should have been more selected more carefully. I could say it was the lack of knowledge because of which I missed a most important part of looking at all the variables in the Data set and check whether all the variables are in same measurements or not. Because some of them are Boolean, some are in mcg/L which led to a problem when working on the model.

Were there any variables you felt could have helped in the analysis?

There are so many other variables which we can include that are more connected to heart failure than that we have in the data set; smoking, gender, serum sodium do not have an too much of an impact when compared to some other reasons that we have out there. Some of them are stated below

1. An arteriovenous **(AV) fistula** is an abnormal connection between an artery and a vein in which blood flows directly from an artery into a vein, bypassing some capillaries. An arteriovenous (AV) fistula is an abnormal connection between an artery and a vein.
2. **Sepsis** occurs when chemicals released in the bloodstream to fight an infection trigger inflammation throughout the body. This can cause a cascade of changes that damage multiple

organ systems, leading them to fail, sometimes even resulting in death. Symptoms include fever, difficulty breathing, low blood pressure, fast heart rate, and mental confusion. Treatment includes antibiotics and intravenous fluids.

3. Coronary artery disease (CAD), also called coronary heart disease (CHD), **ischemic heart disease** (IHD), or simply heart disease, involves the reduction of blood flow to the heart muscle due to build-up of plaque (atherosclerosis) in the arteries of the heart. It is the most common of the cardiovascular diseases.

Were there any assumptions made you felt were incorrect?

Yes, I felt there will not be much difference in variable data measurement as we can run the model with only two or three variables that can give us the good model to fit it as there will be a point where the accuracy of the model doesn't change after a certain point (This assumption was purely based on the book knowledge that I had in this semester). So, did not worry about the measurements of all the variables. May be I should have normalized those variables get some accurate results.

What challenges did you face, what did you not fully understand?

- I also took 520 subject in this semester, where we used R for running the models and validating the results. I felt R-Studio was much more user friendly and easier to use than Jupyter Notebooks.
- Some of the topics like survival analysis, time series analysis were too complex to understand from the textbook, it did not cover everything to make us understand the theories. We had to google them which was time consuming.