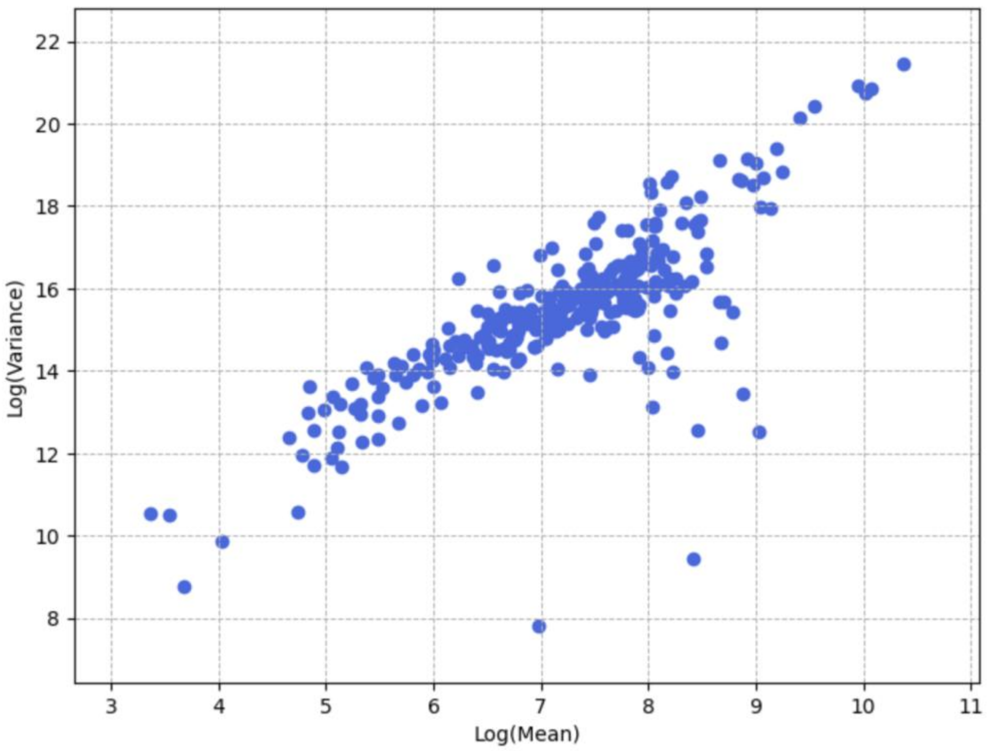
MSCA31010: Linear & Non-Linear Models Winter Quarter 2023

Assignment 5

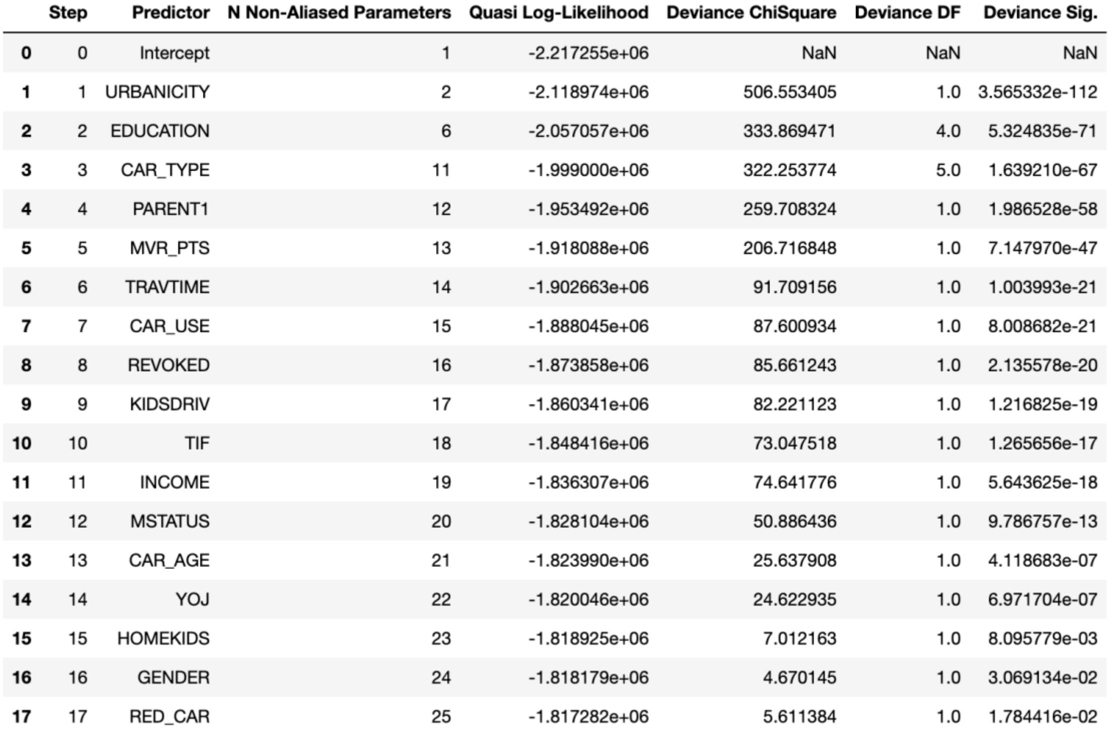
Question 1 (50 points)

a) (10 points). We will first estimate the Tweedie distribution's Power parameter p and Scale parameter ϕ. To this end, we will calculate the sample means and the sample variances of the claim amount for each value combination of the categorical predictors. Then, we will train a linear regression model to help us estimate the two parameters. What are their values? Please provide us with your appropriate chart.



```
Power parameter p : 1.2840608802234919
Scale parameter phi : 495.39032035257253
```

b) (10 points). We will use the Forward Selection method to enter predictors into our model. Our entry threshold is 0.05. Please provide a summary report of the Forward Selection in a table. The report should include (1) the Step Number, (2) the Predictor Entered, (3) the Model Degree of Freedom (i.e., the number of non-aliased parameters), (4) the Quasi-Loglikelihood value, (5) the Deviance Chi-squares statistic between the current and the previous models, (6) the corresponding Deviance Degree of Freedom, and (7) the corresponding Chi-square significance.



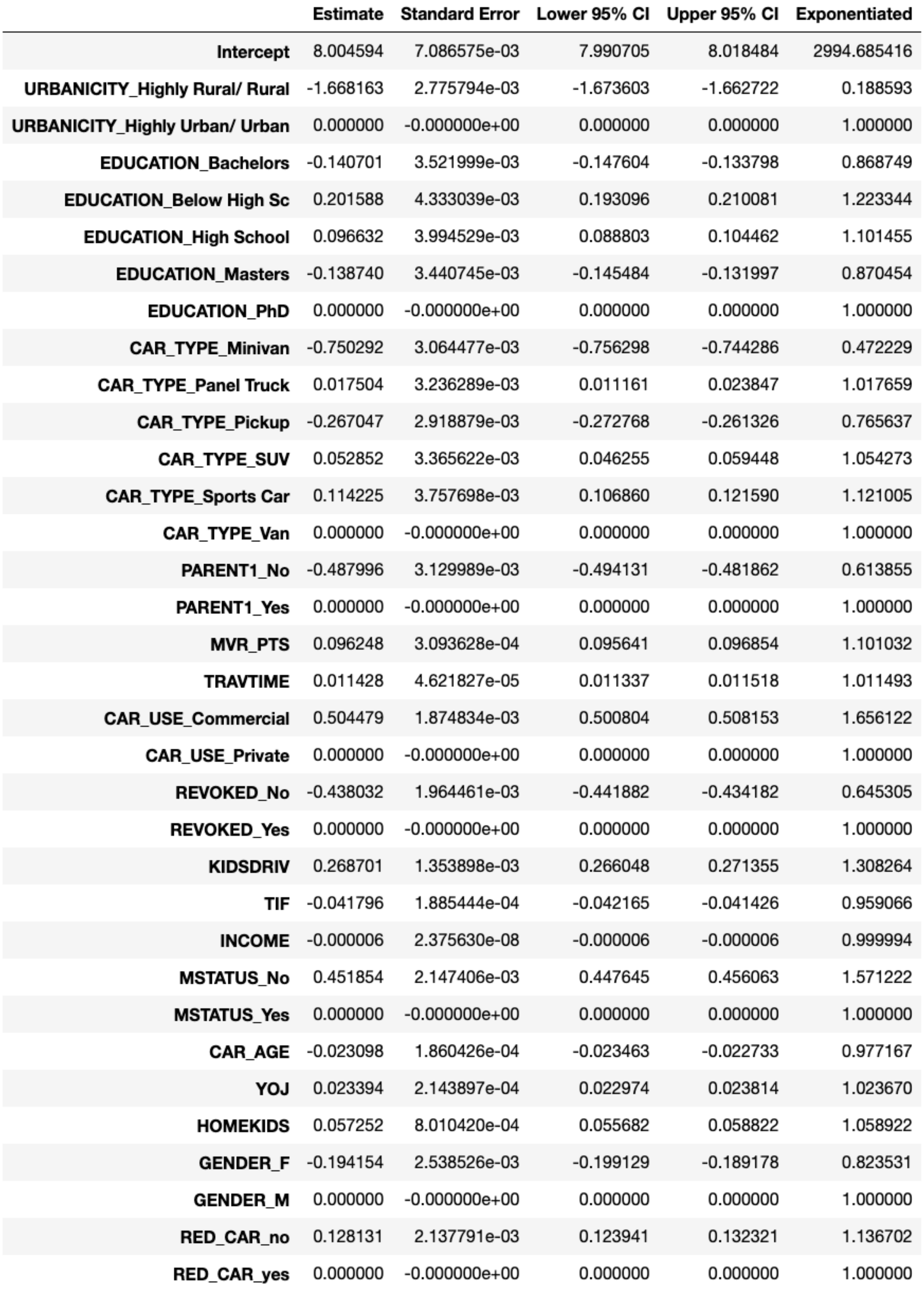
	Step	Predictor	N Non-Aliased Parameters	Quasi Log-Likelihood	Deviance ChiSquare	Deviance DF	Deviance Sig.
0	0	Intercept	1	-2.217255e+06	NaN	NaN	NaN
1	1	URBANICITY	2	-2.118974e+06	506.553405	1.0	3.565332e-112
2	2	EDUCATION	6	-2.057057e+06	333.869471	4.0	5.324835e-71
3	3	CAR_TYPE	11	-1.999000e+06	322.253774	5.0	1.639210e-67
4	4	PARENT1	12	-1.953492e+06	259.708324	1.0	1.986528e-58
5	5	MVR_PTS	13	-1.918088e+06	206.716848	1.0	7.147970e-47
6	6	TRAVTIME	14	-1.902663e+06	91.709156	1.0	1.003993e-21
7	7	CAR_USE	15	-1.888045e+06	87.600934	1.0	8.008682e-21
8	8	REVOKED	16	-1.873858e+06	85.661243	1.0	2.135578e-20
9	9	KIDSDRIV	17	-1.860341e+06	82.221123	1.0	1.216825e-19
10	10	TIF	18	-1.848416e+06	73.047518	1.0	1.265656e-17
11	11	INCOME	19	-1.836307e+06	74.641776	1.0	5.643625e-18
12	12	MSTATUS	20	-1.828104e+06	50.886436	1.0	9.786757e-13
13	13	CAR_AGE	21	-1.823990e+06	25.637908	1.0	4.118683e-07
14	14	YOJ	22	-1.820046e+06	24.622935	1.0	6.971704e-07
15	15	HOMEKIDS	23	-1.818925e+06	7.012163	1.0	8.095779e-03
16	16	GENDER	24	-1.818179e+06	4.670145	1.0	3.069134e-02
17	17	RED_CAR	25	-1.817282e+06	5.611384	1.0	1.784416e-02

c) (10 points). We will calculate the Root Mean Squared Error, the Relative Error, the Pearson correlation, and the Distance correlation between the observed and the predicted claim amounts of your final model. Please comment on their values.

```
RMSE : 4116.064009419275
Relative Error : 1.0078985075249884
Pearson Correlation : 0.18768098705727354
Distance Correlation : 0.27019055675583464
```

A lower RMSE indicates better performance of the model. In this case, the RMSE value of 4116.06 suggests that there is a significant difference between the predicted and actual values. The relative error value of 1.0079 indicates that the predicted values are slightly higher than the actual values. The Pearson correlation value of 0.1876 suggests that there is a weak positive correlation between the predicted and actual values. The distance correlation value of 0.2702 indicates a weak dependence between the predicted and actual values**. Overall, these values suggest that the predictive model may not be performing very well. The RMSE value is quite high, and the correlation values suggest weak relationships between the predicted and actual values.**

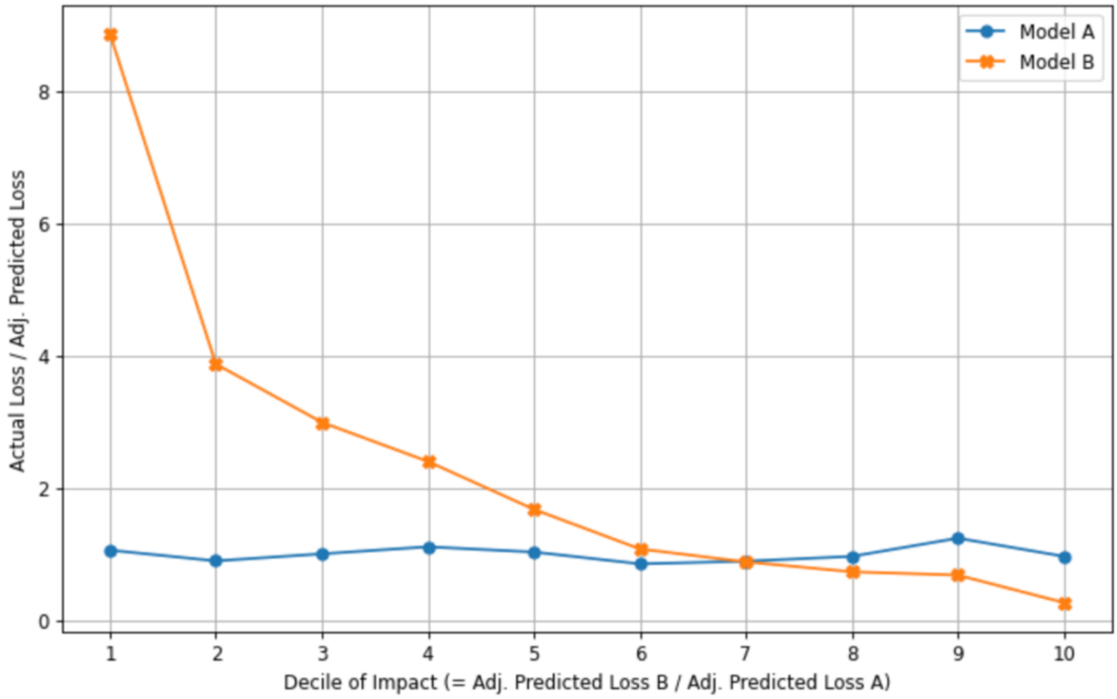
d) (10 points). Please show a table of the complete set of parameters of your final model (including the aliased parameters). Besides the parameter estimates, please also include the standard errors, the 95% asymptotic confidence intervals, and the exponentiated parameter estimates. Conventionally, aliased parameters have zero standard errors and confidence intervals. Please also provide us with the final estimate of the Tweedie distribution's scale parameter ϕ.



	Estimate	Standard Error	Lower 95% CI	Upper 95% CI	Exponentiated
Intercept	8.004594	7.086575e-03	7.990705	8.018484	2994.685416
URBANICITY_Highly Rural/ Rural	-1.668163	2.775794e-03	-1.673603	-1.662722	0.188593
URBANICITY_Highly Urban/ Urban	0.000000	-0.000000e+00	0.000000	0.000000	1.000000
EDUCATION_Bachelors	-0.140701	3.521999e-03	-0.147604	-0.133798	0.868749
EDUCATION_Below High Sc	0.201588	4.333039e-03	0.193096	0.210081	1.223344
EDUCATION_High School	0.096632	3.994529e-03	0.088803	0.104462	1.101455
EDUCATION_Masters	-0.138740	3.440745e-03	-0.145484	-0.131997	0.870454
EDUCATION_PhD	0.000000	-0.000000e+00	0.000000	0.000000	1.000000
CAR_TYPE_Minivan	-0.750292	3.064477e-03	-0.756298	-0.744286	0.472229
CAR_TYPE_Panel Truck	0.017504	3.236289e-03	0.011161	0.023847	1.017659
CAR_TYPE_Pickup	-0.267047	2.918879e-03	-0.272768	-0.261326	0.765637
CAR_TYPE_SUV	0.052852	3.365622e-03	0.046255	0.059448	1.054273
CAR_TYPE_Sports Car	0.114225	3.757698e-03	0.106860	0.121590	1.121005
CAR_TYPE_Van	0.000000	-0.000000e+00	0.000000	0.000000	1.000000
PARENT1_No	-0.487996	3.129989e-03	-0.494131	-0.481862	0.613855
PARENT1_Yes	0.000000	-0.000000e+00	0.000000	0.000000	1.000000
MVR_PTS	0.096248	3.093628e-04	0.095641	0.096854	1.101032
TRAVTIME	0.011428	4.621827e-05	0.011337	0.011518	1.011493
CAR_USE_Commercial	0.504479	1.874834e-03	0.500804	0.508153	1.656122
CAR_USE_Private	0.000000	-0.000000e+00	0.000000	0.000000	1.000000
REVOKED_No	-0.438032	1.964461e-03	-0.441882	-0.434182	0.645305
REVOKED_Yes	0.000000	-0.000000e+00	0.000000	0.000000	1.000000
KIDSDRIV	0.268701	1.353898e-03	0.266048	0.271355	1.308264
TIF	-0.041796	1.885444e-04	-0.042165	-0.041426	0.959066
INCOME	-0.000006	2.375630e-08	-0.000006	-0.000006	0.999994
MSTATUS_No	0.451854	2.147406e-03	0.447645	0.456063	1.571222
MSTATUS_Yes	0.000000	-0.000000e+00	0.000000	0.000000	1.000000
CAR_AGE	-0.023098	1.860426e-04	-0.023463	-0.022733	0.977167
YOJ	0.023394	2.143897e-04	0.022974	0.023814	1.023670
HOMEKIDS	0.057252	8.010420e-04	0.055682	0.058822	1.058922
GENDER_F	-0.194154	2.538526e-03	-0.199129	-0.189178	0.823531
GENDER_M	0.000000	-0.000000e+00	0.000000	0.000000	1.000000
RED_CAR_no	0.128131	2.137791e-03	0.123941	0.132321	1.136702
RED_CAR_yes	0.000000	-0.000000e+00	0.000000	0.000000	1.000000

Final estimate of Tweedie distribution's scale parameter $\phi = \mathbf{319.3817800403747}$

e) (10 points). Please generate a Two-way Lift chart for comparing your final model with the Intercept only model. Based on the chart, what will you conclude about your final model?



From the chart we can infer that, our final model performs best when we target the top 10% of the population (decile of Impact = 1) but gradually decreases as the target population increases. Our model performance is very poor as we can see that post decile of impact = 7 our model performance is worse that our baseline model, i.e., the Intercept only model.

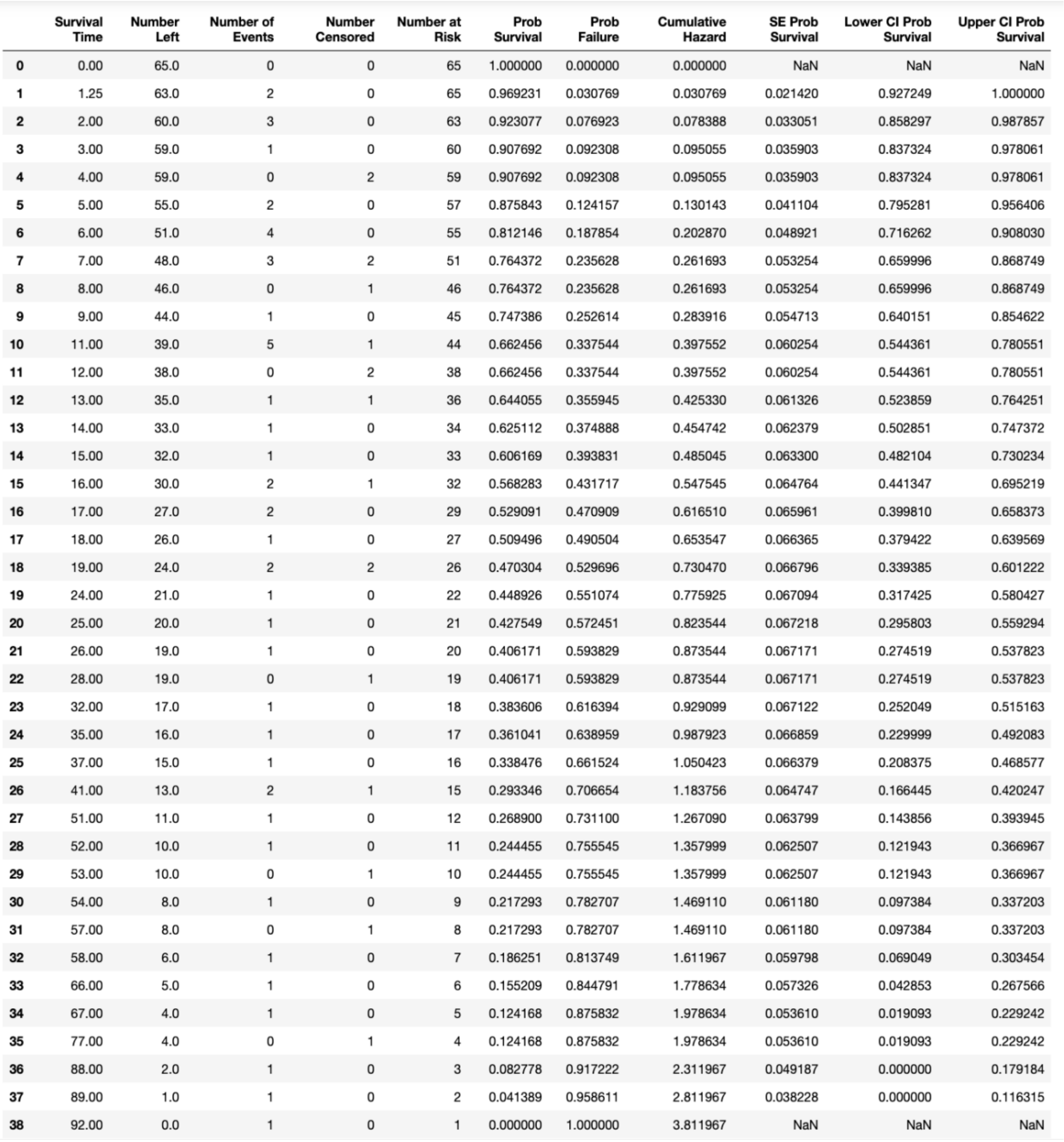
Question 2 (50 points)

a) (10 points). How many risk sets are there?

A risk set is defined as the set of individuals who are at risk of experiencing an event at a given time point. For example, in our case, the risk set at a particular time point consists of all patients who have not yet died in our multiple myeloma study.

From the life table we can identify that the number of risk sets is **38**

b) (10 points). We will use the Kaplan-Meier Product Limit Estimator to create the life table. Please provide us with the life table.



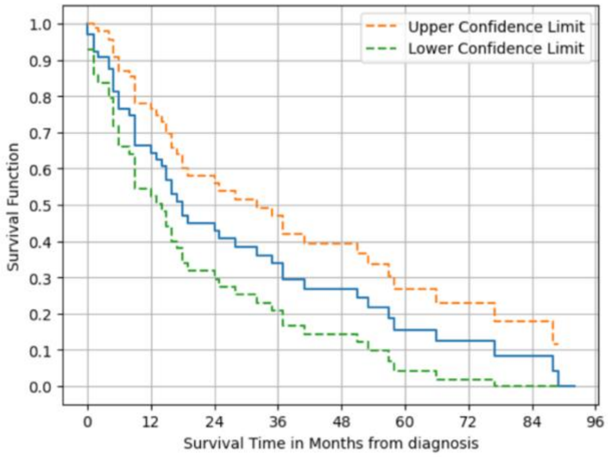
	Survival Time	Number Left	Number of Events	Number Censored	Number at Risk	Prob Survival	Prob Failure	Cumulative Hazard	SE Prob Survival	Lower CI Prob Survival	Upper CI Prob Survival
0	0.00	65.0	0	0	65	1.000000	0.000000	0.000000	NaN	NaN	NaN
1	1.25	63.0	2	0	65	0.969231	0.030769	0.030769	0.021420	0.927249	1.000000
2	2.00	60.0	3	0	63	0.923077	0.076923	0.078388	0.033051	0.858297	0.987857
3	3.00	59.0	1	0	60	0.907692	0.092308	0.095055	0.035903	0.837324	0.978061
4	4.00	59.0	0	2	59	0.907692	0.092308	0.095055	0.035903	0.837324	0.978061
5	5.00	55.0	2	0	57	0.875843	0.124157	0.130143	0.041104	0.795281	0.956406
6	6.00	51.0	4	0	55	0.812146	0.187854	0.202870	0.048921	0.716262	0.908030
7	7.00	48.0	3	2	51	0.764372	0.235628	0.261693	0.053254	0.659996	0.868749
8	8.00	46.0	0	1	46	0.764372	0.235628	0.261693	0.053254	0.659996	0.868749
9	9.00	44.0	1	0	45	0.747386	0.252614	0.283916	0.054713	0.640151	0.854622
10	11.00	39.0	5	1	44	0.662456	0.337544	0.397552	0.060254	0.544361	0.780551
11	12.00	38.0	0	2	38	0.662456	0.337544	0.397552	0.060254	0.544361	0.780551
12	13.00	35.0	1	1	36	0.644055	0.355945	0.425330	0.061326	0.523859	0.764251
13	14.00	33.0	1	0	34	0.625112	0.374888	0.454742	0.062379	0.502851	0.747372
14	15.00	32.0	1	0	33	0.606169	0.393831	0.485045	0.063300	0.482104	0.730234
15	16.00	30.0	2	1	32	0.568283	0.431717	0.547545	0.064764	0.441347	0.695219
16	17.00	27.0	2	0	29	0.529091	0.470909	0.616510	0.065961	0.399810	0.658373
17	18.00	26.0	1	0	27	0.509496	0.490504	0.653547	0.066365	0.379422	0.639569
18	19.00	24.0	2	2	26	0.470304	0.529696	0.730470	0.066796	0.339385	0.601222
19	24.00	21.0	1	0	22	0.448926	0.551074	0.775925	0.067094	0.317425	0.580427
20	25.00	20.0	1	0	21	0.427549	0.572451	0.823544	0.067218	0.295803	0.559294
21	26.00	19.0	1	0	20	0.406171	0.593829	0.873544	0.067171	0.274519	0.537823
22	28.00	19.0	0	1	19	0.406171	0.593829	0.873544	0.067171	0.274519	0.537823
23	32.00	17.0	1	0	18	0.383606	0.616394	0.929099	0.067122	0.252049	0.515163
24	35.00	16.0	1	0	17	0.361041	0.638959	0.987923	0.066859	0.229999	0.492083
25	37.00	15.0	1	0	16	0.338476	0.661524	1.050423	0.066379	0.208375	0.468577
26	41.00	13.0	2	1	15	0.293346	0.706654	1.183756	0.064747	0.166445	0.420247
27	51.00	11.0	1	0	12	0.268900	0.731100	1.267090	0.063799	0.143856	0.393945
28	52.00	10.0	1	0	11	0.244455	0.755545	1.357999	0.062507	0.121943	0.366967
29	53.00	10.0	0	1	10	0.244455	0.755545	1.357999	0.062507	0.121943	0.366967
30	54.00	8.0	1	0	9	0.217293	0.782707	1.469110	0.061180	0.097384	0.337203
31	57.00	8.0	0	1	8	0.217293	0.782707	1.469110	0.061180	0.097384	0.337203
32	58.00	6.0	1	0	7	0.186251	0.813749	1.611967	0.059798	0.069049	0.303454
33	66.00	5.0	1	0	6	0.155209	0.844791	1.778634	0.057326	0.042853	0.267566
34	67.00	4.0	1	0	5	0.124168	0.875832	1.978634	0.053610	0.019093	0.229242
35	77.00	4.0	0	1	4	0.124168	0.875832	1.978634	0.053610	0.019093	0.229242
36	88.00	2.0	1	0	3	0.082778	0.917222	2.311967	0.049187	0.000000	0.179184
37	89.00	1.0	1	0	2	0.041389	0.958611	2.811967	0.038228	0.000000	0.116315
38	92.00	0.0	1	0	1	0.000000	1.000000	3.811967	NaN	NaN	NaN

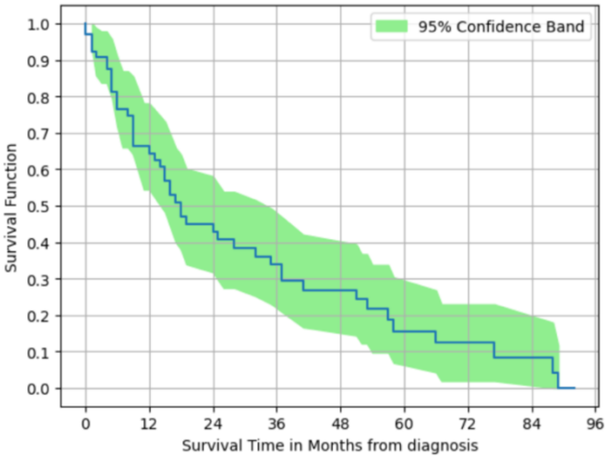
c) (10 points). According to the life table, what is the Probability of Survival and the Cumulative Hazard at a survival time of 18 months? What do these two values mean to a layperson?

Survival Time : 18 months
Probability of Survival : 0.509496
Cumulative Hazard : 0.653547

These values tell us that tell us that the probability of surviving up to 18 months is **50.95%**, and the cumulative probability of experiencing the event of interest (death due to multiple myeloma in our case) up to this time point (i.e., 18 months) is **65.35%**

d) (10 points). Please generate the Survival Function graph using the Kaplan-Meier Product Limit Estimator life table. Since we measure the Time variable in the number of months, we will specify the x-axis ticks from 0 with an increment of 12. Besides plotting the Survival Function versus Time, you must also add the 95% Confidence Band. You might use the matplotlib fill_between() function to generate the Confidence Band as a band around the Survival Function. To receive the full credits, you must label the chart elements properly.





e) (10 points). Use Linear Interpolation to determine the Median Survival Time (in number of months) from the Kaplan-Meier Product Limit Estimator life table. Please round your answer up to the tenths place.

```
Median Survival Time (in number of months) is 18.2 months.
```