Aryan Sharma

Lung Cancer

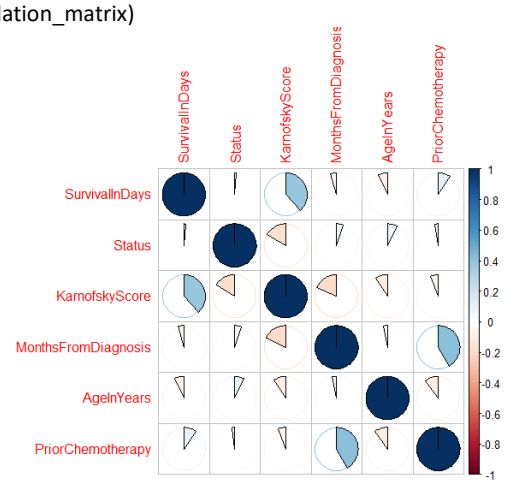
Exploratory Data Analysis

The data has 8 variables and 137 observations. The data is based on patients who are dealing with lung cancer and their treatment. Based on the data the patients are treated in two ways A) Chemotherapy & B) Chemotherapy with the introduction of a new medicine drug

69 people are treated with the standard treatment of chemotherapy and 68 people are treated with the new test treatment of drug plus chemotherapy. Out of the 137 patients, 9 are still alive and the rest 128 have died due to cancer. Out of the 137 patients, 40 patients have gone with chemotherapy prior.

Correlation:

```
correlation_matrix <- cor(data[c("SurvivalInDays",
"Status","KarnofskyScore","MonthsFromDiagnosis","AgeInYears","PriorChemotherapy")])
corrplot(correlation_matrix, method = "pie")
                print(correlation_matrix)
```

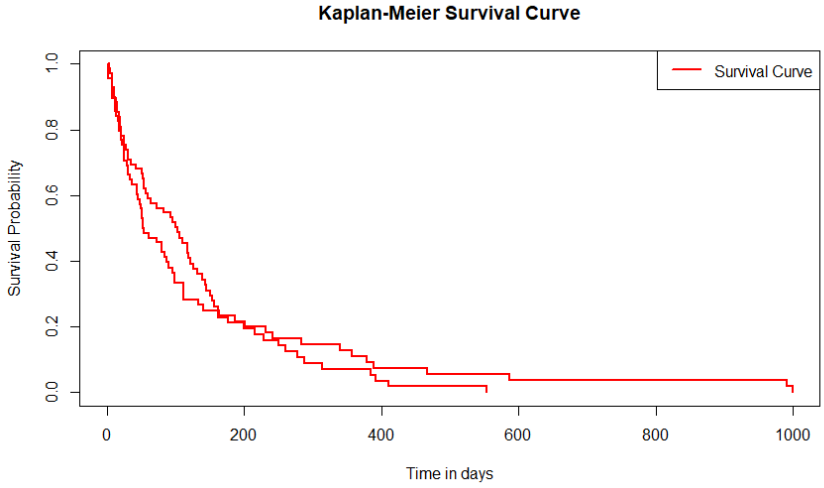


There is no correlation between the variables

Kalplan-Meier Analysis:

Kaplan-Meier analysis is a descriptive analysis hence it is not a predictive model and not explanatory.

```
fit <- survfit(Surv(SurvivalInDays,Status) ~ Treatment, data = data)
plot(fit, main = "Kaplan-Meier Survival Curve", xlab = "Time in days", ylab = "Survival Probability",col
="Red",lwd=2,)
legend("topright", legend = "Survival Curve", col = "Red", lty = 1, lwd = 2)
```



As the data is grouped by treatment, the curve which ends at 500 is treatment 1 i.e. chemotherapy and the other line is treatment 2, i.e. chemo plus the drug. From the graph it can be observed for the initial treatment the results are the same in both the treatments and then the chances of survival are more for treatment 2.

Lung Cancer Patient probability of Survival for 1 year(365 days):

```
time_point <- 365
summary_fit <- summary(fit, times = time_point)
print(summary_fit$surv)
```

```
> print(summary_fit$surv)
[1] 0.07080893 0.10977353
```

Survival probability after 1 year for treatment1 that is only chemotherapy is 70.8% and for treatment 2 that is Chemotherapy is 10.97%

Lung Cancer Patient probability of Survival for 6 months(180 days):

```
time_point <- 180
summary_fit <- summary(fit, times = time_point)
print(summary_fit$surv)
```

```
> print(summary_fit$surv)
[1] 0.2124268 0.2328529
```

Survival probability after 6Months for treatment1 that is only chemotherapy is 21.24% and for treatment 2 that is Chemotherapy is 23.28%

Survival rate summary

```
             n events median 0.95LCL 0.95UCL
treatment=1 69     64  103.0      59     132
treatment=2 68     64   52.5      44      95
```

From the analysis we can say after a lung cancer patient is treated with treatment 1 the median life expectancy is 103 days whereas for treatment 2 which is still in the testing phase the life expectancy median is 52 days. This means treatment 1 is better based on the limited data we had but treatment 2 is still in the test stage as they're testing the drug with chemotherapy treatment.

Predictor table:

Predictor	Effect	Rationale
DV: Survival		
Treatment	+/-	Any treatment can be better as the treatment2 is still in test phase
CellType	+	Celltype defines the type of cancer. Cancer is unaccounted growth of the cells
AgeinYear	+	Age of the patient can lead to less survival rate as the immune system of the body is weakened as we age
MonthsFromDiagnosis	+/-	If a patient got diagnostic prior or in early stage the chances of survival may increase
PriorChemotherapy	+	This shows whether the patient has been going for previous chemotherapy, going for the treatment might kill the cancer cell and increase survival rate
Excluded: KarnofskyScore		

Regression Analysis

We have taken 4 models for this analysis, Cox PH model and Parametric survival models. In parameteric survival models I have taken exponential, Weibull and loglogistic models.

```
cox_model=  coxph(Surv(SurvivalInDays,Status) ~ Treatment + CellType + AgeInYears  +
MonthsFromDiagnosis + PriorChemotherapy, data = data)
exp= survreg(Surv(SurvivalInDays,Status) ~ Treatment + CellType + AgeInYears  + MonthsFromDiagnosis +
PriorChemotherapy,dist = "exponential", data = data)
weibull= survreg(Surv(SurvivalInDays,Status) ~ Treatment + CellType + AgeInYears  +
MonthsFromDiagnosis + PriorChemotherapy,dist = "weibull", data = data)
loglogistic= survreg(Surv(SurvivalInDays,Status) ~ Treatment + CellType + AgeInYears  +
MonthsFromDiagnosis + PriorChemotherapy,dist = "loglogistic", data = data)
```

	Dependent variable:			
	SurvivalInDays			
	Cox prop. hazards	exponential	Weibull	survreg: loglogistic
	(1)	(2)	(3)	(4)
Treatment	0.167 (0.202)	-0.136 (0.199)	-0.135 (0.202)	-0.210 (0.225)
CellType2	1.040*** (0.278)	-1.081*** (0.264)	-1.081*** (0.268)	-1.076*** (0.298)
CellType3	1.188*** (0.299)	-1.233*** (0.274)	-1.230*** (0.279)	-0.939*** (0.339)
CellType4	0.290 (0.285)	-0.308 (0.276)	-0.303 (0.282)	0.091 (0.327)
AgeInYears	0.005 (0.010)	-0.007 (0.010)	-0.007 (0.010)	0.001 (0.011)
MonthsFromDiagnosis	0.010 (0.009)	-0.010 (0.009)	-0.010 (0.009)	-0.010 (0.013)
PriorChemotherapy	-0.007 (0.023)	0.010 (0.023)	0.009 (0.023)	-0.020 (0.027)
Constant		6.099*** (0.619)	6.087*** (0.631)	5.145*** (0.743)
Observations	137	137	137	137
R2	0.180			
Max. Possible R2	0.999			
Log Likelihood	-491.890	-733.085	-733.046	-738.397
chi2 (df = 7)		36.272***	30.091***	23.737***
Wald Test	26.510*** (df = 7)			
LR Test	27.117*** (df = 7)			
Score (Logrank) Test	28.073*** (df = 7)			

Interpretation

1. **Treatment:** Exponential, Weibull and loglogistic models show Treatment has a negative effect on Survival rate whereas the Cox Ph model apprehends that the Treatment 2 would have a positive effect on the survival rate. (Where treatment2 is Chemotherapy and drug) CoxPh model: 167% increase, Exponential: 136% decrease, Weibull 135% decrease, Loglogistic: 210% decrease.
2. **CellType:** Celltype1 as the base that is squamous. Exponential, Weibull and loglogistic models show celltype2, celltype3 and celltype4 has a negative effect on Survival rate whereas the Cox Ph model apprehends that the celltype2, celltype3 and celltype4 would have a positive effect on the survival rate.
3. **AgeInYears:** Age in Years is the age of the patient who is going through the treatment. According to Cox model as the age increases the chances of survival increases by 5%, whereas in exponential the chances decline by 7%, Weibull : decline by 7%, whereas according to loglogistic the chances of survival increase by 1%.
4. **MonthsFromDiagnosis:** This variable reflects the number of months since the patient has been diagnosed with cancer. According to exponential, Weibull and loglogistic model as the month increases it has a negative effect of 10% on survival rate of the patient while coxph model says a 10% increase in survival rate.
5. **PriorChemotherapy:** Patients who are previously treated with Chemo have a positive effect on survival probability based on exponential and Weibull model and loglogistic and Coxph model apprehends that prior chemotherapy has a negative effect on their survival.

Recommendations:

Conduct sensitivity analysis to assess the robustness of your findings to different modeling assumptions, variable specifications, or analytical methods. This involves varying key parameters and assessing their impact on model results.