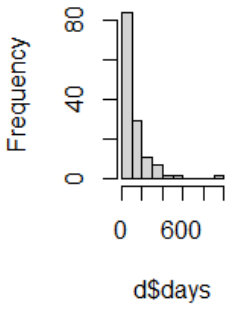
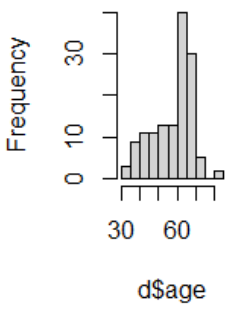
Lung Cancer

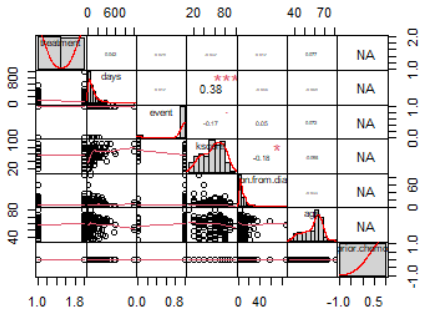
1. The transformations necessary for this dataset are:

→Changed the variables for ease of understanding:
prior chemo values to 1,0 from 10,0
cell types from numeric to str factors

2. The visualizations shown below are histograms and correlation plot. Histogram of age shows that higher age people are more who have cancer. There doesn't seem to be any noticeable correlation between variables.



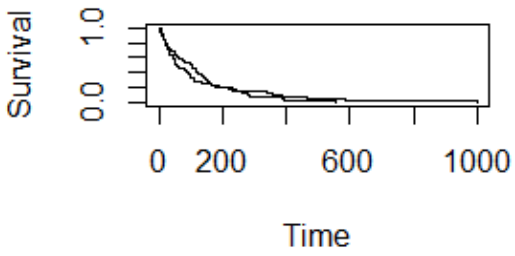




3. Predictor Table:

Predictor	Effect	Rationale
DV: Days+Event		
treatment	?	To answer the questions and interpret the results of standard chemo vs chemo with new drug.
celltype	?	Categorical data, need to check which type will have more chances of survival.
kscore	+	If a patient performs more ordinary tasks better, his survival chances should be more.
Month from diagnosis	-	More the time from diagnosis, less would be the survival chances logically.
age	-	More the age, more health risks and comorbidities might lead to less survival chances.
Prior chemo	+/-	It's the basic therapy for cancer and it can surely affect the time to death.

Kaplan-Meier survival graphs for patients with the test vs standard treatment



Seeing the KM curve, we can say that, with respect to data, 1% survival chance for standard group is 13.5 months whereas, for treatment group with new drug is 32 months.

Probability that the patient will survive for 6 months (183 days) and 1 year (365 days) on the standard treatment vs the test treatment

	6 months	1 year
Standard	20-21%	5-6%
Treatment	21.6-22%	9-10.5%

The values in the above answer are referenced from the below snap of results of kaplan meier survival model. For convinience only the resultset with 183days and 365days have been taken.

d$treatment=1

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
177	14	1	0.2124	0.0514	0.13218	0.341
200	12	1	0.1947	0.0501	0.11761	0.322
216	11	1	0.1770	0.0486	0.10340	0.303
228	10	1	0.1593	0.0468	0.08956	0.283
250	9	1	0.1416	0.0448	0.07614	0.263
260	8	1	0.1239	0.0426	0.06318	0.243
278	7	1	0.1062	0.0400	0.05076	0.222
287	6	1	0.0885	0.0371	0.03896	0.201
314	5	1	0.0708	0.0336	0.02793	0.180
384	4	1	0.0531	0.0295	0.01788	0.158

d$treatment=2

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
164	15	1	0.2329	0.0529	0.14920	0.363
186	14	1	0.2162	0.0517	0.13538	0.345
201	13	1	0.1996	0.0503	0.12181	0.327
231	12	1	0.1830	0.0488	0.10851	0.308
242	10	1	0.1647	0.0472	0.09389	0.289
283	9	1	0.1464	0.0454	0.07973	0.269
340	8	1	0.1281	0.0432	0.06609	0.248
357	7	1	0.1098	0.0407	0.05304	0.227
378	6	1	0.0915	0.0378	0.04067	0.206

Median number of days where a patient can be expected to survive if they are on the standard vs the test treatment:

On considering the survival probability > 0.50, Patient can be expected to survive for 100 days for standard treatment and 51 days for treatment with a new drug.

Three semi-parametric and parametric models to estimate the marginal effects:

```
cox <- coxph(y ~ treatment + celltype + kscore + mon.from.diag + age + prior.chemo, data = d)
exp <- survreg(y ~ treatment + celltype + kscore + mon.from.diag + age + prior.chemo, data = d,
dist="exponential")
weibull <- survreg(y ~ treatment + celltype + kscore + mon.from.diag + age + prior.chemo, data = d,
dist="weibull")
```

```
                  exp(coef) exp(-coef) lower .95 upper .95
treatment2           1.3426     0.7448    0.8939    2.0166
celltypeSmall Cell   2.3669     0.4225    1.3799    4.0597
celltypeadeno        3.3071     0.3024    1.8336    5.9647
celltypelarge        1.4938     0.6695    0.8583    2.5996
kscore               0.9677     1.0334    0.9573    0.9782
mon.from.diag        1.0001     0.9999    0.9823    1.0182
age                  0.9913     1.0087    0.9734    1.0096
prior.chemo          1.0742     0.9309    0.6813    1.6937

Concordance= 0.736  (se = 0.021 )
Likelihood ratio test= 62.1  on 8 df,   p=2e-10
Wald test            = 62.37  on 8 df,   p=2e-10
Score (logrank) test = 66.74  on 8 df,   p=2e-11
```

Interpretations based on coxPH semi-parametric model.

→Exp(coef) defines the hazard ratio which is the ratio of risk of event happening to risk of event not happening. Here the event is death. Therefore, we want HR <1.
→Confidently we can only say the effects of few variables who remain >1 or <1 even in 95% intervals, which are celltypes (squamous, small, adeno), kscore.
→Celltypes: People with small cell are 2.3 times more likely to die within the given time period as compared to cell type squamous keeping other parameters constant.
Similarly, people with adeno cell are 3.3 times more likely to die within the given time period as compared to cell type squamous keeping other parameters constant.
Increased Karnofsky score would decrease the risk of death by 4%.
→The treatment group with the new drug is 34% more likely to die(risk of occurrence of event) as compared to the standard group. Since the confidence intervals lie between <1 and >1, we can't be sure of the exact effect related to the new drug.
→Similarly for the age variable, increase in age would 1% reduction in risk of death, which doesn't make lot of sense. The confidence interval lies in both region >1, <1, hence we can't be sure of the effect.
→Treatment group with prior chemo will have increased risk of death by 7% which can't be stated firmly because according to confidence interval, considering lower limit, it will reduce the risk of death by 32%, while upper limit suggests that it will increase the risk of death by 69%. This is a very wide interval to say correctly about the effect of prior chemo.

→Stargazer

	Dependent variable:		
	y		
	Cox prop. hazards	exponential	Weibull
	(1)	(2)	(3)
treatment2	0.295 (0.208)	-0.220 (0.199)	-0.229 (0.187)
celltypeSmall Cell	0.862*** (0.275)	-0.820*** (0.262)	-0.826*** (0.246)
celltypeadeno	1.196*** (0.301)	-1.113*** (0.276)	-1.133*** (0.258)
celltypelarge	0.401 (0.283)	-0.377 (0.273)	-0.398 (0.255)
kscore	-0.033*** (0.006)	0.031*** (0.005)	0.030*** (0.005)
mon.from.diag	0.0001 (0.009)	-0.0003 (0.009)	-0.0005 (0.008)
age	-0.009 (0.009)	0.006 (0.009)	0.006 (0.009)
prior.chemo	0.072 (0.232)	-0.049 (0.227)	-0.044 (0.212)
Constant		3.189*** (0.704)	3.262*** (0.663)
Observations	137	137	137
R2	0.364		
Max. Possible R2	0.999		
Log Likelihood	-474.397	-716.159	-715.551
chi2 (df = 8)		70.124***	65.080***
Wald Test	62.370*** (df = 8)		
LR Test	62.104*** (df = 8)		
Score (Logrank) Test	66.737*** (df = 8)		