EXPLORATORY_ANALYSIS

PRAHLAD

Management And Field Engineers Are Looking For Answers From Analytics!

- Q1 –> Which Turbine had most number of downtimes?
- Q2 –> Which Turbine had maximum downtime? And What was the reason?
- Q3 –> Can we have look at summary stats?Over the period of time which are bad and good performing turbines?
- Q4 –> Is there any underperformance of any turbine? What is the reason and energy loss due to it?
- Q5 –> Are there any anomalies observed in any component?

Read Data

```
library(data.table)
data_operational <-
fread("E:/PROJECTS/E/Data_Sample_Operational_Data.csv",sep=";",stringsAsFacto
rs =FALSE)
data_alarms <-
fread("E:/PROJECTS/E/Data_Sample_Alarms.csv",sep=";",stringsAsFactors =FALSE)
#powercurve <- read.csv("pc.csv")
```

Understanding Operational Data

Renaming Missing Columns And Converting to Readable Formats

```
library(lubridate)

##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:data.table':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday,
##     week, yday, year

## The following object is masked from 'package:base':
##
##     date

library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:lubridate':
##
##     intersect, setdiff, union

## The following objects are masked from 'package:data.table':
##
##     between, first, last

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(magrittr)
#Rename columns#
setnames(data_operational,"(No column name)","TIMESTAMP")
#Convert to R readable timestamp
data_operational$TIMESTAMP <-
substr(data_operational$TIMESTAMP,start=1,stop=19)
data_operational$TIMESTAMP <-as.POSIXct(data_operational$TIMESTAMP,format
="%Y-%m-%d %H:%M:%S")
data_operational$YEAR <-year(data_operational$TIMESTAMP)
data_operational$MONTH <-month(data_operational$TIMESTAMP)
data_operational$YEAR_MONTH <-
paste(data_operational$YEAR,data_operational$MONTH,sep="_")
#Remove Duplicates If Found
data_operational <-unique(data_operational,by=c("TIMESTAMP","TurbineName"))
names(data_operational) %<>%toupper()
```

Range Of Data

```
max(data_operational$TIMESTAMP)

## [1] "2017-09-01 IST"

min(data_operational$TIMESTAMP)

## [1] "2016-09-01 IST"
```

One-year Of Data Provided

How Many Turbines In The Farm?

```
unique(data_operational$TURBINENAME)
```

```
##  [1] "ADTA0500" "ADTA0600" "ADTA0700" "ADTA0900" "ADTA1000" "ADTA1100"
##  [7] "ADTA1200" "ADTB0400" "ADTB0500" "ADTB0600" "ADTB0700" "ADTB0800"
## [13] "ADTB0900" "ADTB1000" "ADTB1100" "ADTB1200" "ADTC0300" "ADTC0400"
## [19] "ADTC0500" "ADTC0600"
```

Total turbines in the farm are 20

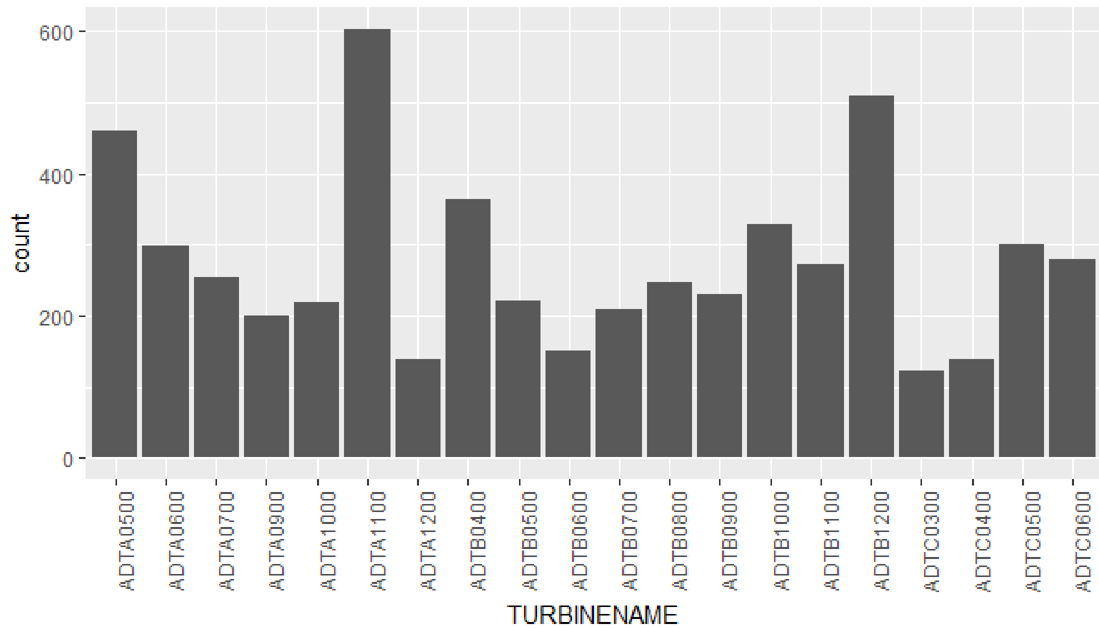
Understanding Events Data

Renaming Missing Columns And Converting to Readable Formats

```
#Convert to r readable timestamp#
data_alarms$TimeDetected <-substr(data_alarms$TimeDetected,start=1,stop=19)
data_alarms$TimeDetected <-as.POSIXct(data_alarms$TimeDetected,format ="%Y-
%m-%d %H:%M:%S")
data_alarms$TimeReset <-substr(data_alarms$TimeReset,start=1,stop=19)
data_alarms$TimeReset <-as.POSIXct(data_alarms$TimeReset,format ="%Y-%m-%d
%H:%M:%S")
data_alarms$YEAR <-year(data_alarms$TimeReset)
data_alarms$MONTH <-month(data_alarms$TimeReset)
data_alarms$YEAR_MONTH <-paste(data_alarms$YEAR,data_alarms$MONTH,sep="_")
#Calculate breakdown hours#
data_alarms$BREAKDOWNHOURS <-
difftime(data_alarms$TimeReset,data_alarms$TimeDetected,units="mins")
data_alarms$BREAKDOWNHOURS <-round(as.numeric(data_alarms$BREAKDOWNHOURS),0)
names(data_alarms) %<>%toupper()
```

Q1- Which Turbine Has Most Number Of Breakdowns?

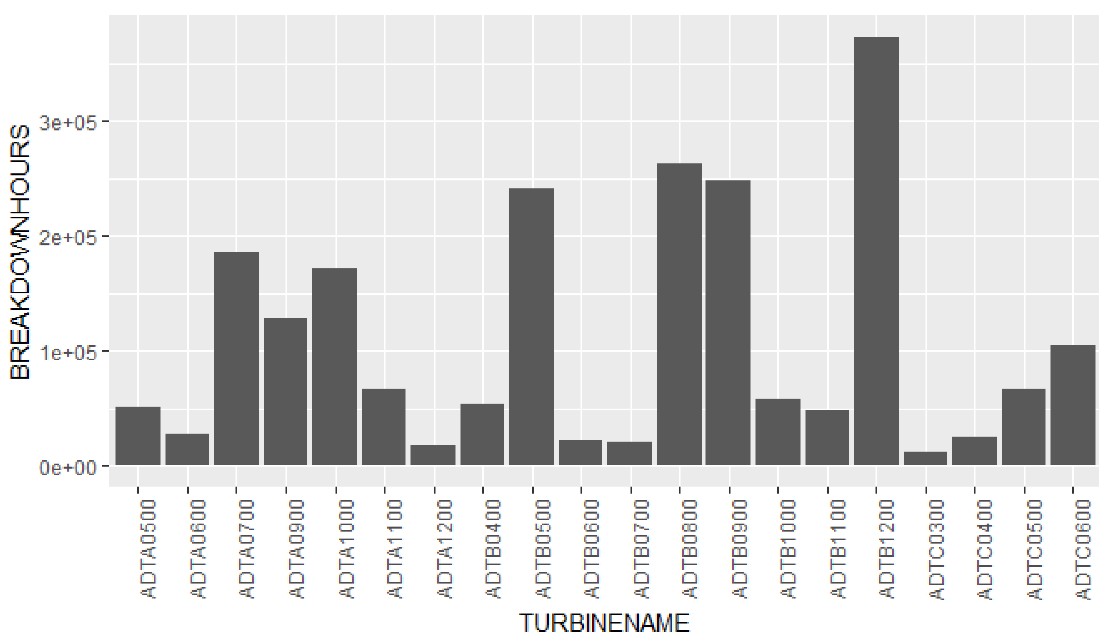
```
library(ggplot2)
qplot(data=data_alarms,x=TURBINENAME)+geom_bar()+theme(axis.text.x
=element_text(angle =90, hjust =1))
```



It is seen that Turbine-"ADTA1100" has most number of breakdowns

Q2- Which Turbine Has Maximum Downtime?

```
ggplot(data=data_alarms,aes(x=TURBINENAME,y=BREAKDOWNHOURS))+geom_bar(stat="i
dentity") +theme(axis.text.x=element_text(angle=90,hjust=1))
```



It is seen that Turbine-"ADTB1200" has maximum downtime

Which Alarm Is Causing For Most Downtime in Turbine-“ADTB1200”?

```
S <-data_alarms %>%filter(TURBINENAME=="ADTB1200")
%>%group_by(TURBINENAME,YEAR,MONTH,ALARMDESCRIPTION,BREAKDOWNHOURS)
%>%summarise(EVENTS_COUNT =n())

arrange(S,desc(BREAKDOWNHOURS))

## # A tibble: 470 x 6
## # Groups:   TURBINENAME, YEAR, MONTH, ALARMDESCRIPTION [244]
##    TURBINENAME  YEAR MONTH ALARMDESCRIPTION    BREAKDOWNHOURS EVENTS_COUNT
##    <chr><dbl><dbl><chr><dbl><int>
##  1 ADTB1200     2017     3 DiffPressHigh 50Â°~          15176            1
##  2 ADTB1200     2016    12 DiffPressHigh 52Â°~          14589            1
##  3 ADTB1200     2017     3 DiffPressHigh 51Â°~          14506            1
##  4 ADTB1200     2017     4 DiffPressHigh 53Â°~          14488            1
##  5 ADTB1200     2017     4 DiffPressHigh 47Â°~          14446            1
##  6 ADTB1200     2017     3 DiffPressHigh 51Â°~          14435            1
##  7 ADTB1200     2017     2 DiffPressHigh 44Â°~          14422            1
##  8 ADTB1200     2017     1 DiffPressHigh 54Â°~          14416            1
##  9 ADTB1200     2017     4 DiffPressHigh 54Â°~          14409            1
## 10 ADTB1200     2017     1 DiffPressHigh 54Â°~          14406            1
## # ... with 460 more rows
```

It is seen that Event-“DiffPressHigh” event attributes maximum to downtime

Q3 –> Summary Stats Of Each Turbine Month-On-Month

Aggregating Alarms Data

```
data_alarms_agg <-data_alarms
%>%select(TURBINENAME,ALARMDESCRIPTION,YEAR_MONTH,BREAKDOWNHOURS)
%>%group_by(TURBINENAME,YEAR_MONTH) %>%summarise(EVENTS_COUNT
=n(),TOTAL_BREAKDOWNHOURS =sum(BREAKDOWNHOURS,na.rm=TRUE))
```

Aggregating Operational Data

```
data_operational$PROD_LATESTAVG_ACTPWRGEN1 <-
as.numeric(data_operational$PROD_LATESTAVG_ACTPWRGEN1)

data_operational$AMB_WINDSPEED_AVG <-
as.numeric(data_operational$AMB_WINDSPEED_AVG)

data_operational_agg <-data_operational
%>%select(TURBINENAME,YEAR_MONTH,PROD_LATESTAVG_ACTPWRGEN1,AMB_WINDSPEED_AVG)
%>%group_by(TURBINENAME,YEAR_MONTH) %>%summarise(PROD_ACTIVEPOWER_AVG
=mean(PROD_LATESTAVG_ACTPWRGEN1,na.rm=TRUE),AMB_WINDSPEED_MONTH_AVG=mean(AMB_
WINDSPEED_AVG,na.rm=TRUE),COUNT_OPS_OBS =n())
```

```r
#Calculate  Total Number Of Observations
# I have taken 31 days as a proxy
data_operational_agg$TOTAL_OPS_OBS <-31*144

#Calculate data_quality
data_operational_agg$DATA_QUALITY <-
round((data_operational_agg$COUNT_OPS_OBS/data_operational_agg$TOTAL_OPS_OBS)
*100,0)
```

Merge Operational Data With Events Data Aggregated

```r
data_merge <-
merge(data_operational_agg,data_alarms_agg,by=c("YEAR_MONTH","TURBINENAME"),a
ll.x=TRUE)

#Remove NA's as Timenot Available fO Events

data_merge <-na.omit(data_merge)


#Viewing Top 10 Good performing Turbines With Data_Quality G.T 95%

data_summary_good_ten <-data_merge %>%filter(DATA_QUALITY>=95)
%>%arrange(TOTAL_BREAKDOWNHOURS)

data_summary_good_ten <-data_summary_good_ten[1:10,]

library(pander)

panderOptions('table.split.table', Inf)

pander(data_summary_good_ten)
```

YEAR_MONTH	TURBINENAME	PROD_ACTIVEPOWER_AVG	AMB_WINDSPEED_MONTH_AVG	COUNT_OPS_OBS	TOTAL_OPS_OBS	DATA_QUALITY	EVENTS_COUNT	TOTAL_BREAKDOWNHOURS
2016_10	ADTB0900	321786	9.524	4464	4464	100	2	5
2016_11	ADTB0600	326875	9.892	4320	4464	97	4	7
2016_11	ADTC0400	323514	9.537	4320	4464	97	4	16
2016_10	ADTB1100	328394	9.758	4464	4464	100	7	20
2016_10	ADTA1200	349156	10.55	4464	4464	100	2	21

2016_11	ADTA1000	350968	10.13	4320	4464	97	3	26
2016_11	ADTA0900	349350	9.684	4320	4464	97	4	32
2016_10	ADTA0600	345404	10.52	4464	4464	100	4	38
2017_8	ADTA0900	164302	6.421	4464	4464	100	2	42
2017_8	ADTA1100	167725	6.361	4464	4464	100	3	42

```
#Viewing Top 10 Bad performing Turbines With Data_Quality G.T 95%

data_summary_bad_ten <-data_merge %>%filter(DATA_QUALITY>=95)
%>%arrange(desc(TOTAL_BREAKDOWNHOURS))

data_summary_bad_ten <-data_summary_bad_ten[1:10,]

panderOptions('table.split.table', Inf)

pander(data_summary_bad_ten)
```

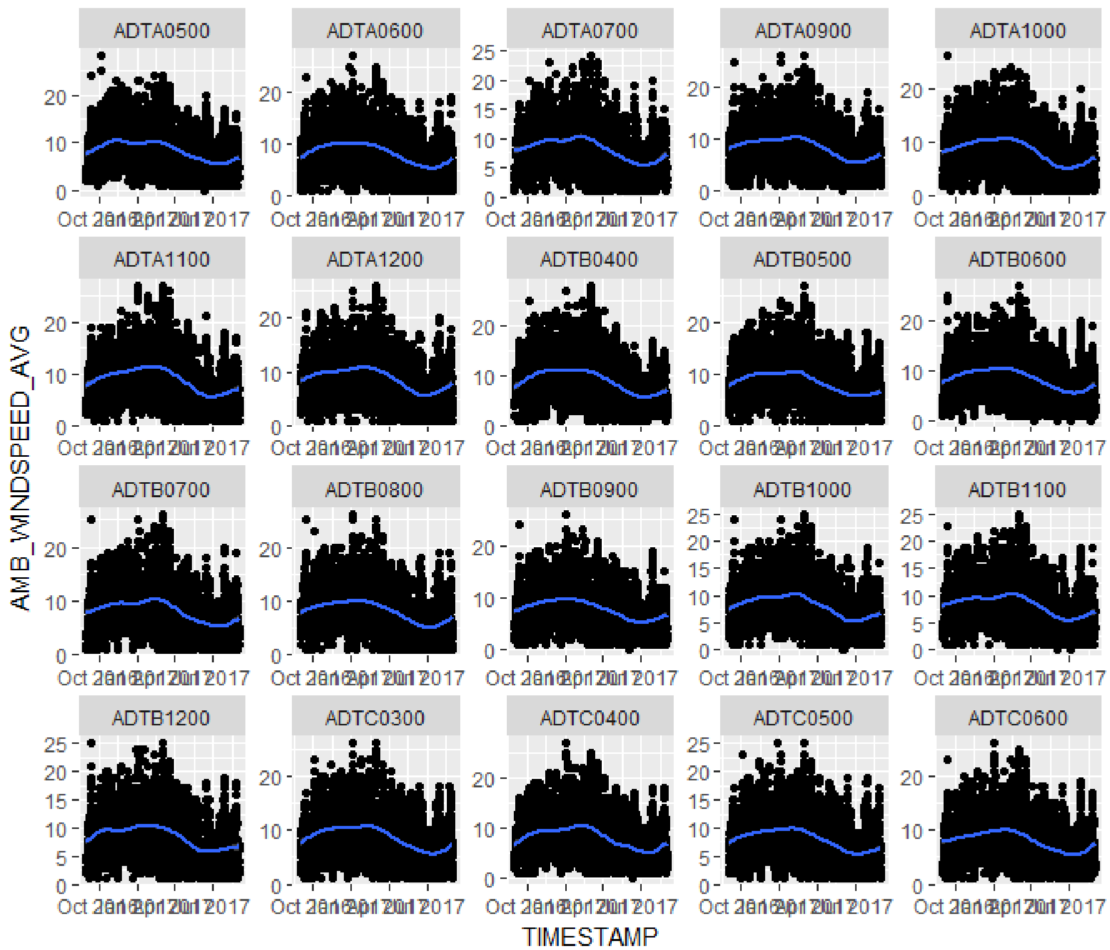
YEAR_MONTH	TURBINENAME	PROD_ACTIVEPOWER_AVG	AMB_WINDSPEED_MONTH_AVG	COUNT_OPS_OBS	TOTAL_OPS_OBS	DATA_QUALITY	EVENTS_COUNT	TOTAL_BREAKDOWNHOURS
2017_1	ADTB0500	333136	10.07	4452	4464	100	13	173191
2016_9	ADTA0700	238561	7.699	4320	4464	97	56	128628
2017_3	ADTB1200	121404	9.965	4462	4464	100	37	128562
2016_11	ADTB0800	333765	9.467	4316	4464	97	27	114150
2017_7	ADTA0900	63203	5.619	4423	4464	99	54	109083
2016_10	ADTB0800	326957	9.92	4464	4464	100	25	68266
2016_10	ADTC0600	306917	9.24	4456	4464	100	6	53086
2017_5	ADTB0900	143298	6.567	4457	4464	100	60	46596
2017_4	ADTB1200	243964	8.316	4279	4464	96	28	45688

2017_5	ADTC0500	165240	6.998	4459	4464	100	73	38002

Q4 –> Exploring Windpattern & Finding Underperformance If Any

Exploring Windpattern

```
ggplot(data=data_operational,aes(x=TIMESTAMP,y=AMB_WINDSPEED_AVG))
+geom_point() +geom_smooth() +facet_wrap(~TURBINENAME,scales="free")
```



It is seen that January,February,March are high wind seasons

Finding Underperformance–Powercurve

```
data_operational$PROD_LATESTAVG_ACTPWRGEN1 <-
as.numeric(data_operational$PROD_LATESTAVG_ACTPWRGEN1)

options(scipen=999)
```

```
ggplot(data=data_operational,aes(x=AMB_WINDSPEED_AVG,y=PROD_LATESTAVG_ACTPWRG
EN1)) +geom_point()+facet_wrap(~TURBINENAME,scales="free")
```

It is seen that there is curtailment on most of the turbines.Also there are some disturbances seen in turbines "ADTC0300","ADTC0400","ADTC0500".Powercurve can be plotted month-on-month to know more details and also we can merge events data to operational data to see if the underperformance data points are due to some events

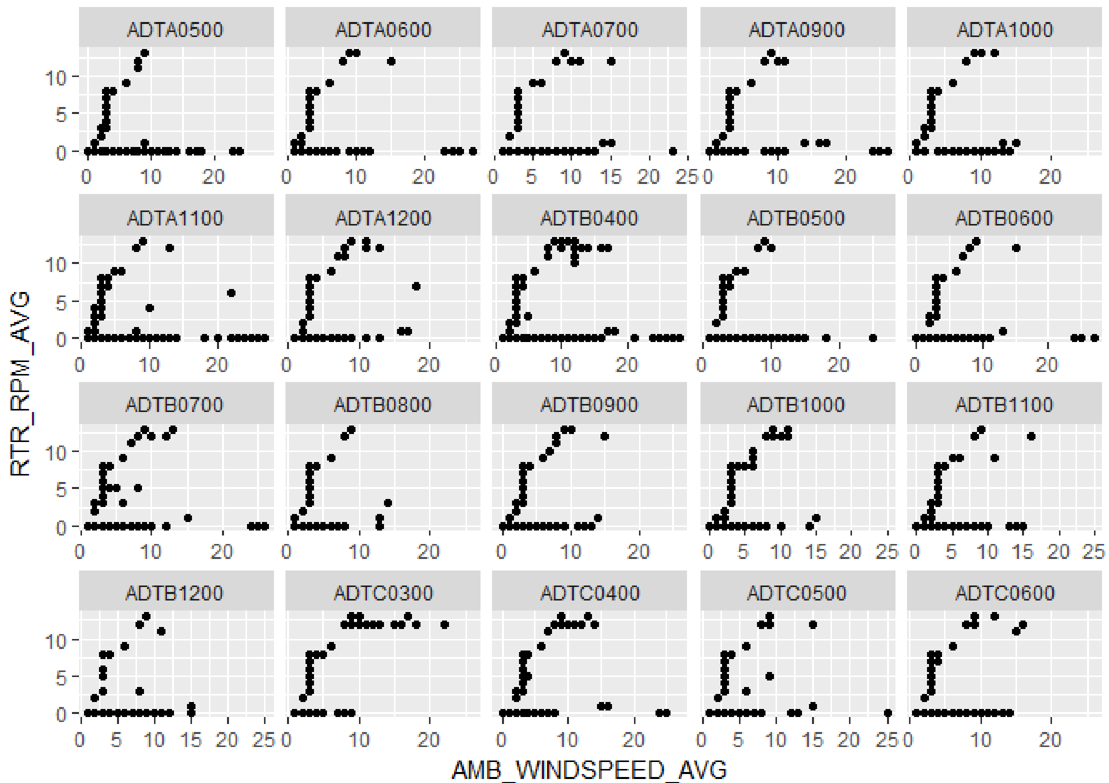
Finding Underperformance–Rotorcurve

```
data_operational$RTR_RPM_AVG  <-as.numeric(data_operational$RTR_RPM_AVG)

## Warning: NAs introduced by coercion

ggplot(data=data_operational,aes(x=AMB_WINDSPEED_AVG,y=RTR_RPM_AVG ))
+geom_point()+facet_wrap(~TURBINENAME,scales="free_x")

## Warning: Removed 1037583 rows containing missing values (geom_point).
```



No clear pattern seen in RotorCurve

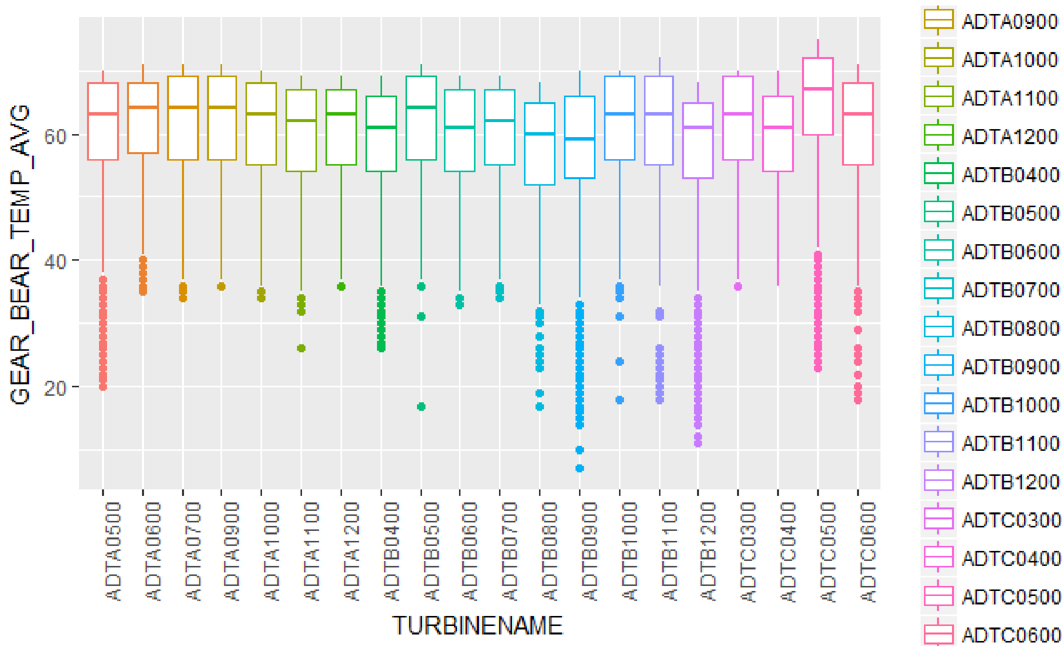
Q5.1 –> Finding Anomalies In Gearbox

Temperature Analysis-Component Considered is Gearbox

Gearbox Bearing Analysis

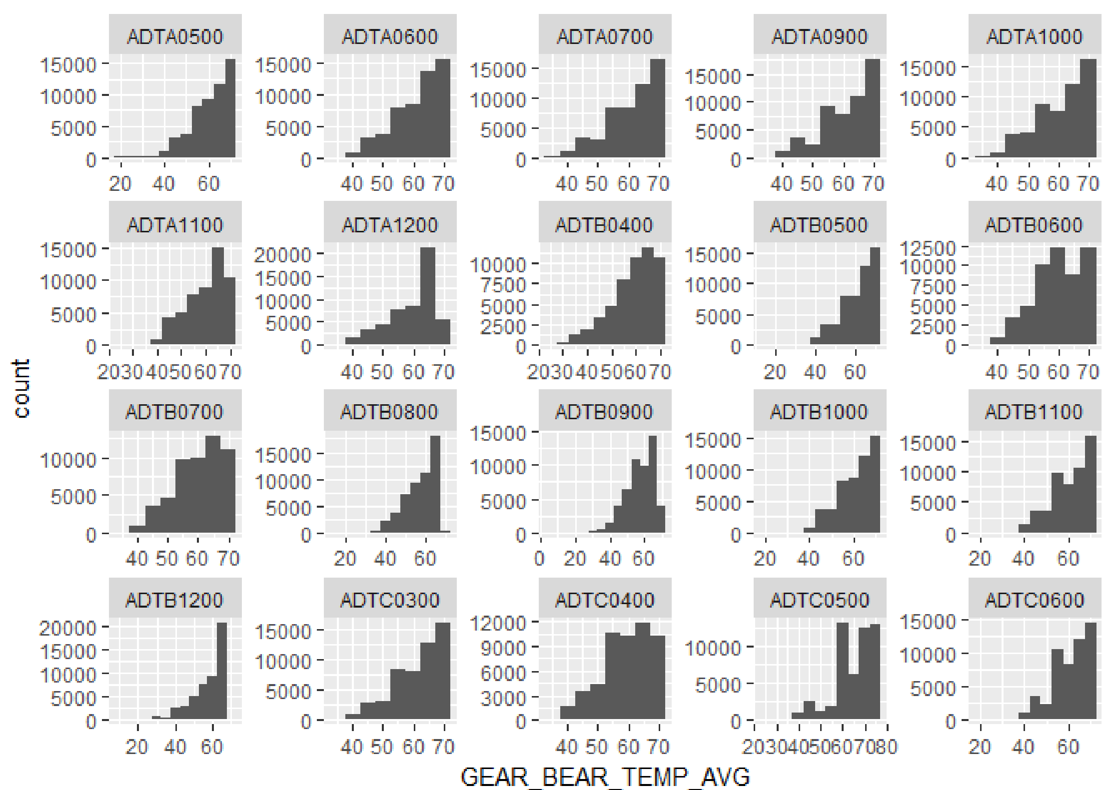
Boxplot of Gearboxbearing for all Turbines

```
ggplot(data=data_operational,aes(x=TURBINENAME,y=GEAR_BEAR_TEMP_AVG,colour=TU
RBINENAME)) +geom_boxplot()
+theme(axis.text.x=element_text(angle=90,hjust=1))
```



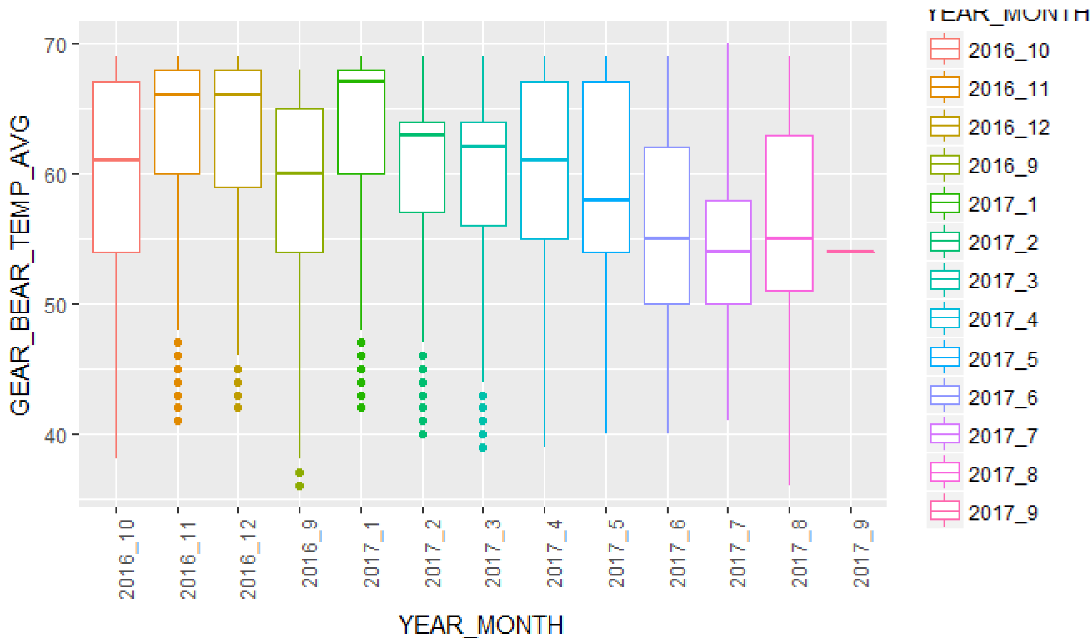
It is seen that gearboxbearing of "ADTC0500" median value is high compared to the rest in the farm

Hist of Gearbearing for all Turbines



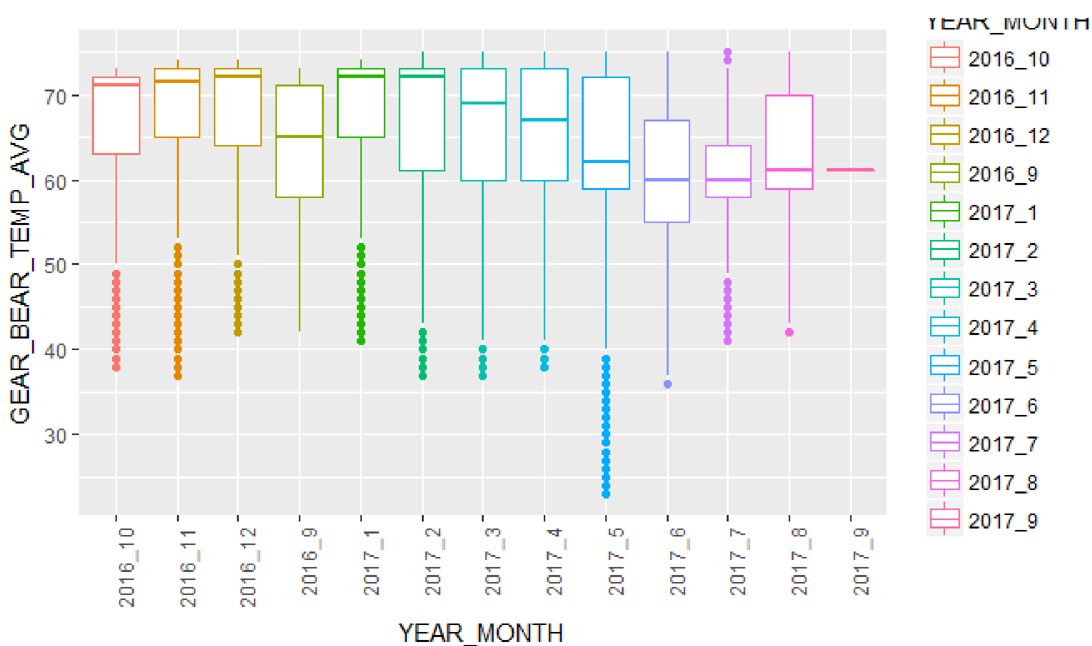
Understanding Healthy Turbine

```
ggplot(data=data_operational[data_operational$TURBINENAME=="ADTC0400",],aes(x
=YEAR_MONTH,y=GEAR_BEAR_TEMP_AVG,colour=YEAR_MONTH))
+geom_boxplot()+theme(axis.text.x=element_text(angle=90,hjust=1))
```



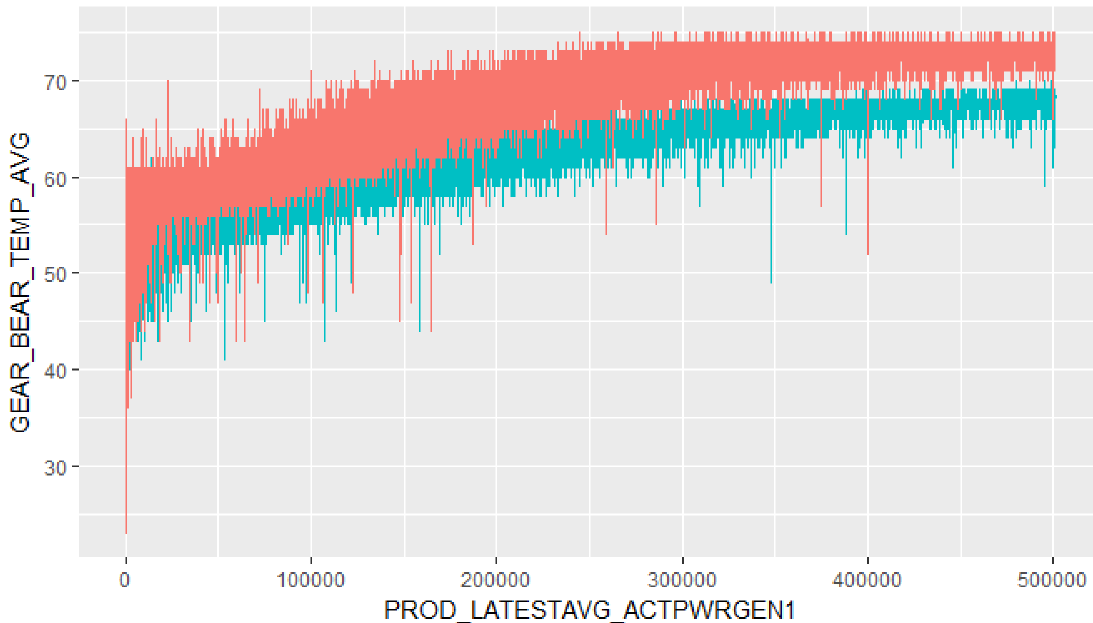
Understanding Anamolous Turbine

```
ggplot(data=data_operational[data_operational$TURBINENAME=="ADTC0500",],aes(x
=YEAR_MONTH,y=GEAR_BEAR_TEMP_AVG,colour=YEAR_MONTH))
+geom_boxplot()+theme(axis.text.x=element_text(angle=90,hjust=1))
```



It is seen that during months of Jan,Feb-2016 the temperature is high reason could also being high wind season.In the month of July-2017 gearbox bearing temp had touched 75 degrees

R/shp Between ActivePower and Gearboxbearing



Correlation between NAC_TEMP_AVG,GEN_RPM_AVG,AMB_TEMP_AVG,RTR_RPM_AVG,PROD_LATESTAVG_ACTPWRGEN1 With Gear_Bear_Temp_Avg for Anamolus Turbine

```
#Convert To Numeric

data_operational$NAC_TEMP_AVG <-as.numeric(data_operational$NAC_TEMP_AVG)
data_operational$GEN_RPM_AVG <-as.numeric(data_operational$GEN_RPM_AVG)

## Warning: NAs introduced by coercion

data_operational$AMB_TEMP_AVG <-as.numeric(data_operational$AMB_TEMP_AVG)
data_operational$RTR_RPM_AVG <-as.numeric(data_operational$RTR_RPM_AVG)
data_operational$PROD_LATESTAVG_ACTPWRGEN1 <-
as.numeric(data_operational$PROD_LATESTAVG_ACTPWRGEN1)

s1 <-subset(data_operational,data_operational$TURBINENAME=="ADTC0500")

s1 <-as.data.frame(s1)

x <-
```

```
s1[c("NAC_TEMP_AVG","GEN_RPM_AVG","AMB_TEMP_AVG","RTR_RPM_AVG","PROD_LATESTAVG_ACTPWRGEN1")]

y <-s1["GEAR_BEAR_TEMP_AVG"]

s3 <-data.frame(x,y)

#Remove Na's##Data Observation Reduces

s3 <-na.omit(s3)

x <-
s3[c("NAC_TEMP_AVG","GEN_RPM_AVG","AMB_TEMP_AVG","RTR_RPM_AVG","PROD_LATESTAVG_ACTPWRGEN1")]

y <-s3["GEAR_BEAR_TEMP_AVG"]

p <-as.data.frame(cor(x,y))

panderOptions('table.split.table', Inf)

pander(p)
```

	GEAR_BEAR_TEMP_AVG
NAC_TEMP_AVG	-0.1087
GEN_RPM_AVG	0.8134
AMB_TEMP_AVG	0.03875
RTR_RPM_AVG	0.8134
PROD_LATESTAVG_ACTPWRGEN1	0.7239

Correlation between NAC_TEMP_AVG,GEN_RPM_AVG,AMB_TEMP_AVG,RTR_RPM_AVG,PROD_LATESTAVG_ACTPWRGEN1 With Gear_Bear_Temp_Avg for Healthy Turbine

```
s2 <-subset(data_operational,data_operational$TURBINENAME=="ADTC0400")

s2 <-as.data.frame(s2)

x <-
s2[c("NAC_TEMP_AVG","GEN_RPM_AVG","AMB_TEMP_AVG","RTR_RPM_AVG","PROD_LATESTAVG_ACTPWRGEN1")]

y <-s2["GEAR_BEAR_TEMP_AVG"]
```

```
s3 <-data.frame(x,y)

#Remove Na's##Data Observation Reduces

s3 <-na.omit(s3)

x <-
s3[c("NAC_TEMP_AVG","GEN_RPM_AVG","AMB_TEMP_AVG","RTR_RPM_AVG","PROD_LATESTAV
G_ACTPWRGEN1")]

y <-s3["GEAR_BEAR_TEMP_AVG"]

p1 <-as.data.frame(cor(x,y))

panderOptions('table.split.table', Inf)

pander(p1)
```

	GEAR_BEAR_TEMP_AVG
NAC_TEMP_AVG	-0.4458
GEN_RPM_AVG	0.9155
AMB_TEMP_AVG	-0.06698
RTR_RPM_AVG	0.9155
PROD_LATESTAVG_ACTPWRGEN1	0.8756

It is seen that there is no difference between two cases ,a scatter plot would be ideal for bivariate analysis to visually see differences

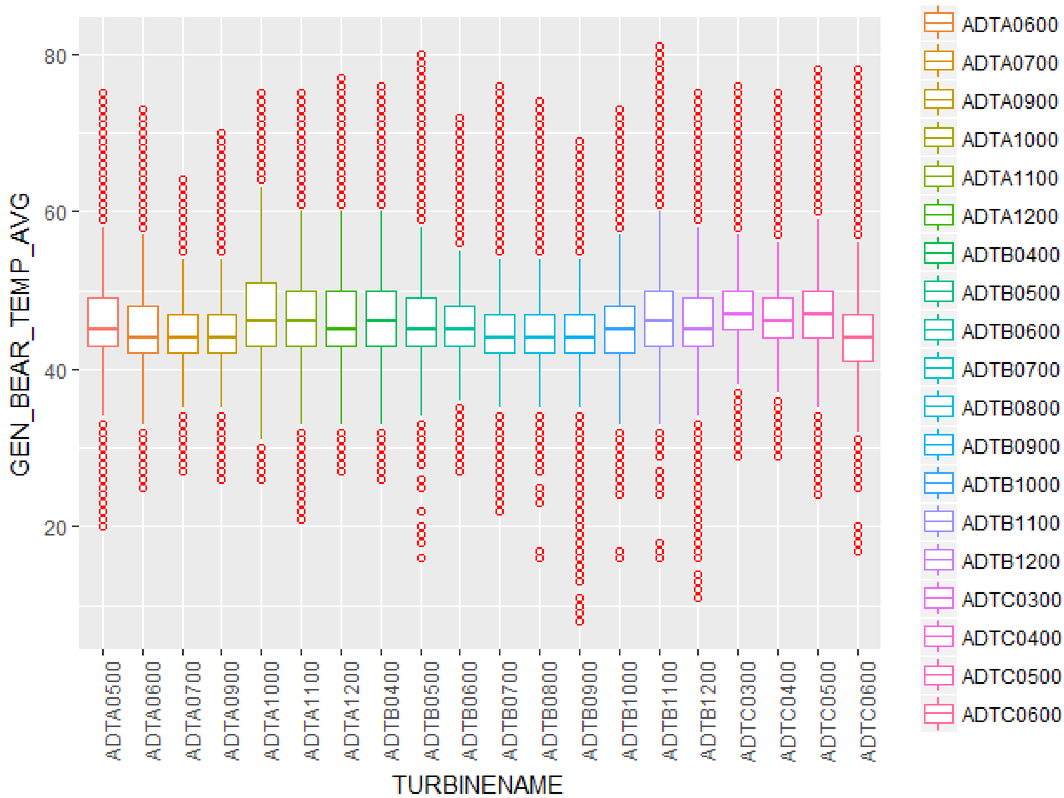
Q5.2 –> Finding Anomalies in Generator

Temperature Analysis-Component Considered is Generator

Generator Bearing1 Analysis

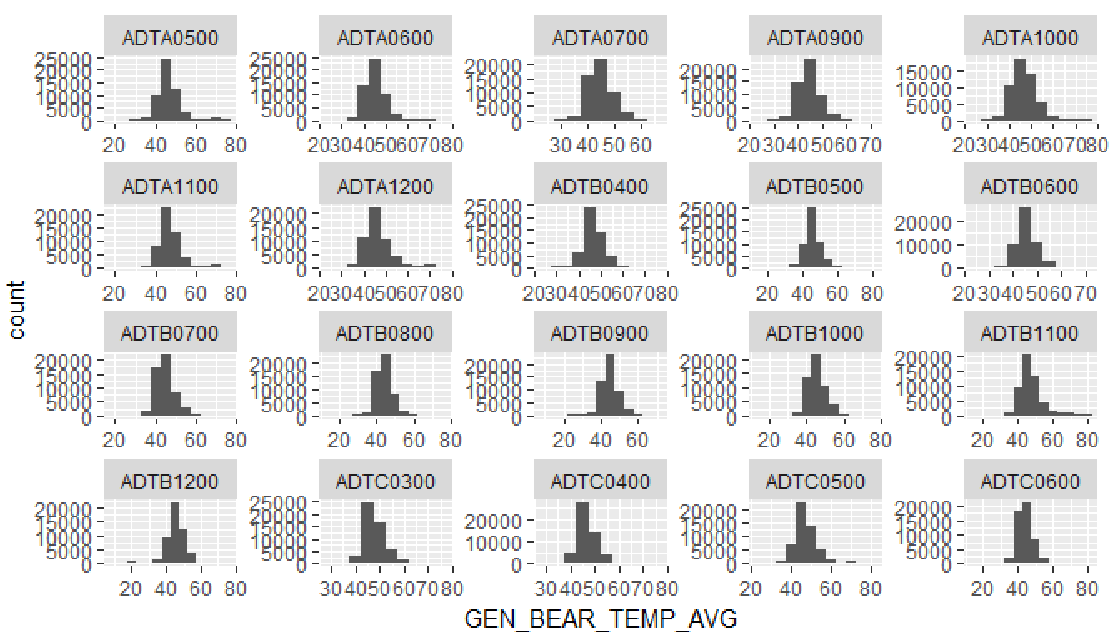
Boxplot of Generator Bearing1 For All Turbines

```
ggplot(data=data_operational,aes(x=TURBINENAME,y=GEN_BEAR_TEMP_AVG,colour=TUR
BINENAME)) +geom_boxplot(outlier.colour ="red", outlier.shape =1)
+theme(axis.text.x=element_text(angle=90,hjust=1))
```



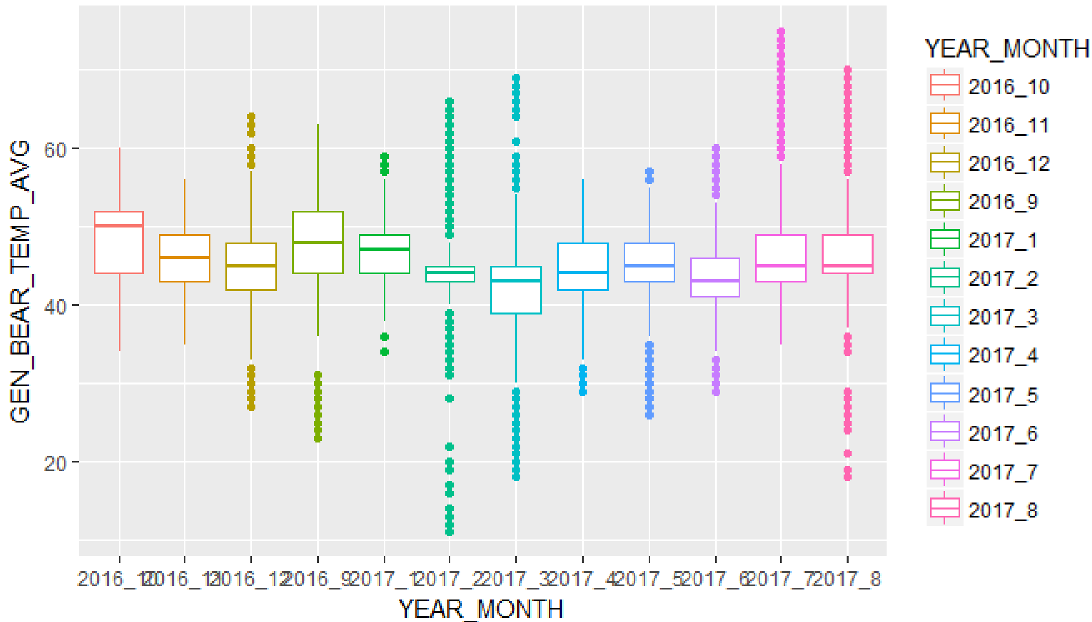
It is seen that generator bearing1 of Turbine-"ADTB1100" has higher temperature data points

Hist of Generatorbearing1 For All Turbines



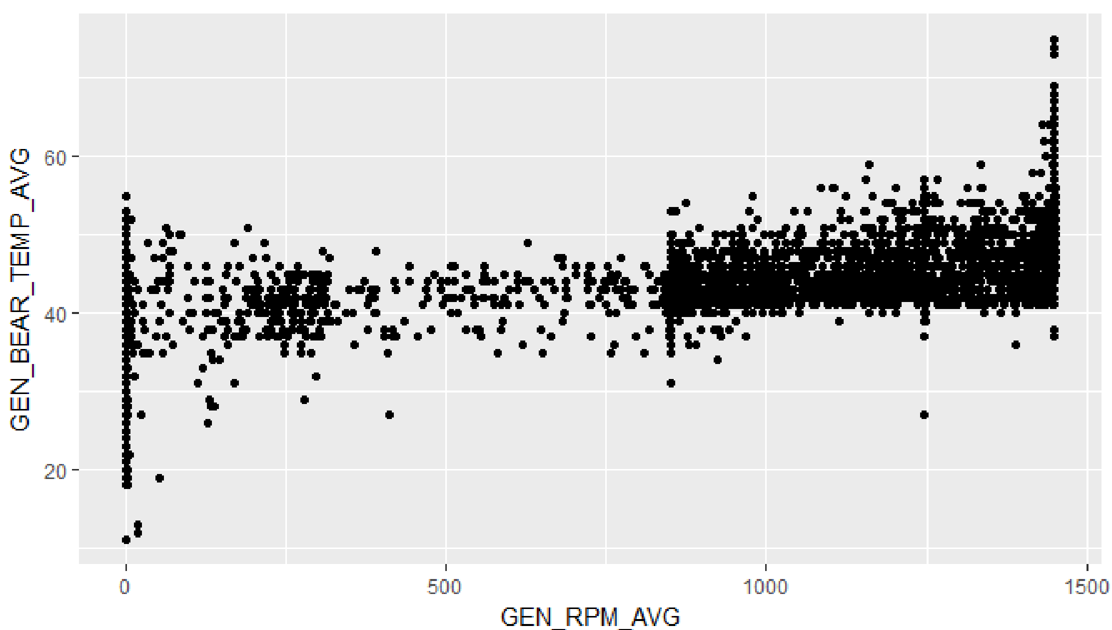
Understanding Healthy Turbine

```
ggplot(data=data_operational[data_operational$TURBINENAME=="ADTB1200",],aes(x
=YEAR_MONTH,y=GEN_BEAR_TEMP_AVG,colour=YEAR_MONTH)) +geom_boxplot()
```



Relationship between Genrpm and Genbearing1 for healthy turbine

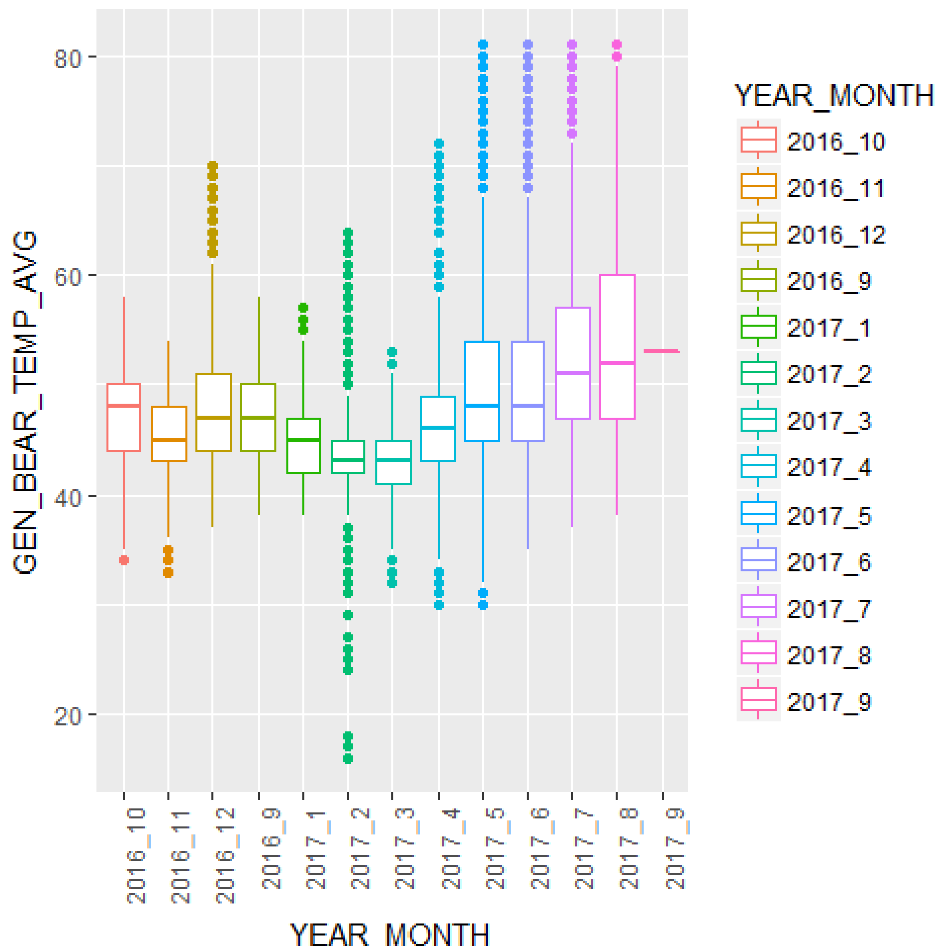
```
ggplot(data=data_operational[data_operational$TURBINENAME=="ADTB1200",],aes(x
=GEN_RPM_AVG,y=GEN_BEAR_TEMP_AVG)) +geom_point()
```



It is seen that as generator rpm increase g.t 1000 rpm, generator bearing1 temperature hasn't increased for "ADTB1200"

Understanding Anamolous Turbine

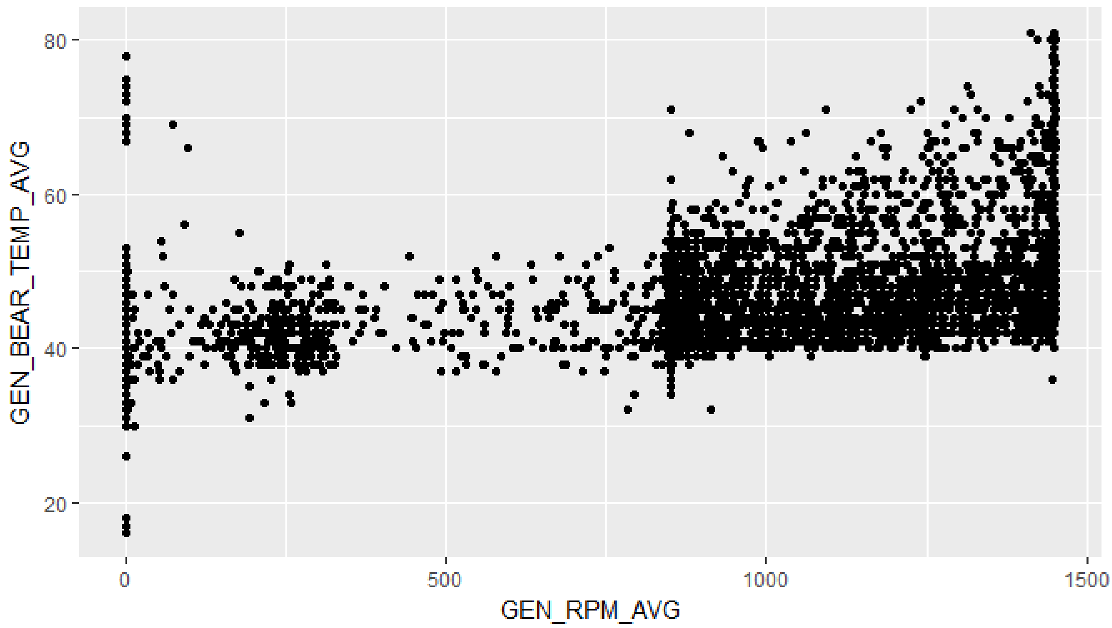
```
ggplot(data=data_operational[data_operational$TURBINENAME=="ADTB1100",],aes(x
=YEAR_MONTH,y=GEN_BEAR_TEMP_AVG,colour=YEAR_MONTH))
+geom_boxplot()+theme(axis.text.x=element_text(angle=90,hjust=1))
```



It is seen that generatorbearing1 of Turbine-"ADTB1100" median temperatures is increasing from April-2017

Relationship between Genrpm and Genbearing1 for Anamolous Turbine

```
ggplot(data=data_operational[data_operational$TURBINENAME=="ADTB1100",],aes(x
=GEN_RPM_AVG,y=GEN_BEAR_TEMP_AVG)) +geom_point()
```



It is seen that as generatore rpm increases g.t 1000 rpm ,generator bearing1 temperature increases

Next Steps !!

-Powercurve can be studied m-o-m and we can calculate energy loss and quantify underperformance.

-Temperature Analysis: More Exploratory analysis for gearbox related parameters should be carried out. Gearboxoil can be studied and there are events associated with it in the alarms data

-Temperature Analysis: Exploratory analysis should be carried out on all the other generator bearings,for example-genbearing1 anomalies was found in "ADTB100" and there can be cases of anomalies in genbearing2.Hence driven end and non driven end shall be studied.

-Once the above two steps are completed, a failure detection model can then be built on each component.