IIF.1105 - IF.2301 - DATA SCIENCE
DATA SCIENCE PROJECT

Principal Component Analysis (PCA) and Linear Regression

FINAL PRESENTATION

62705 GUO Xiaofan

63287 FU Jintao

02/02/2024

Contents

1. Introduction

- **Background**
  - Analysis of Combined Cycle Power Plant Data using PCA and Linear Regression
- **Objective**
  - Explain the aim to analyze the plant's data using Principal Component Analysis and Linear Regression to understand the factors influencing the plant's efficiency.
- **Approach**
  - PCA
  - Linear Regression
- **Importance**
  - Highlight the relevance of this analysis for optimizing power plant operations and enhancing energy efficiency.

2. Data Analysis

2.1 Preliminary Analysis

-Descriptive Statistical Analysis

- **Total Observations**: 9,568
- **Number of Variables**: 5
- **Missing Values**: None detected

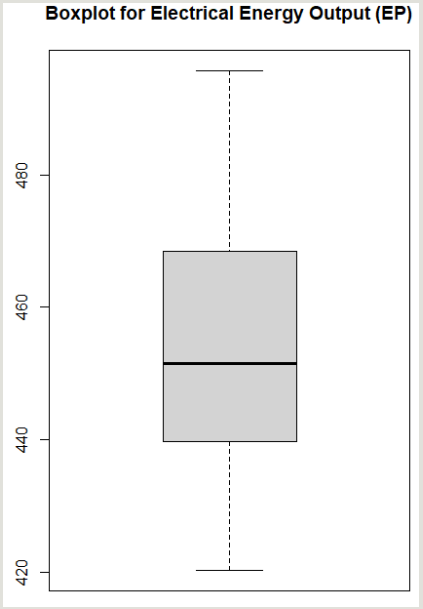
```
> dim(CCPP.data)
[1] 9568    5
> sum(is.na(CCPP.data))
[1] 0
```

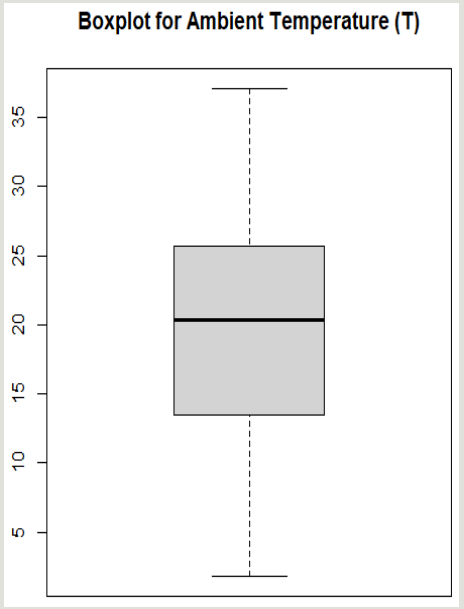
```
> summary(CCPP.data)
       AT               V               AP              RH              EP        
 Min.   : 1.81   Min.   :25.36   Min.   : 992.9   Min.   : 25.56   Min.   :420.3  
 1st Qu.:13.51   1st Qu.:41.74   1st Qu.:1009.1   1st Qu.: 63.33   1st Qu.:439.8  
 Median :20.34   Median :52.08   Median :1012.9   Median : 74.97   Median :451.6  
 Mean   :19.65   Mean   :54.31   Mean   :1013.3   Mean   : 73.31   Mean   :454.4  
 3rd Qu.:25.72   3rd Qu.:66.54   3rd Qu.:1017.3   3rd Qu.: 84.83   3rd Qu.:468.4  
 Max.   :37.11   Max.   :81.56   Max.   :1033.3   Max.   :100.16   Max.   :495.8  
```

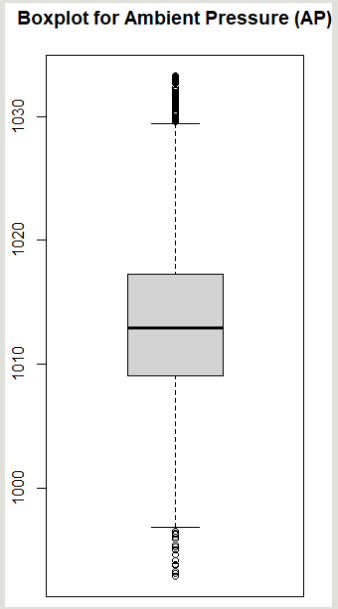
2. Data Analysis

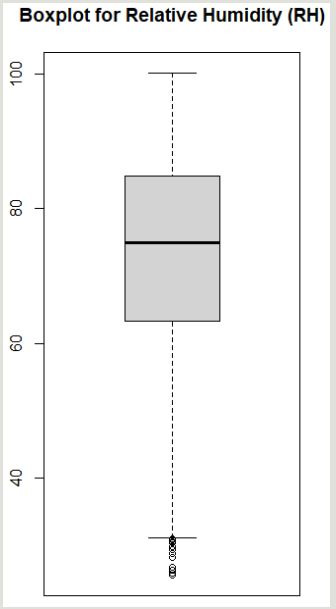
2.1 Preliminary analysis

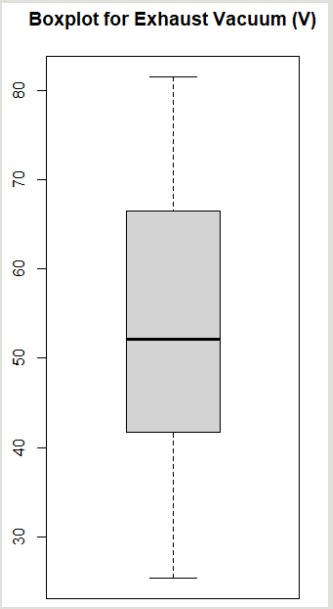
-Box Plot Analysis









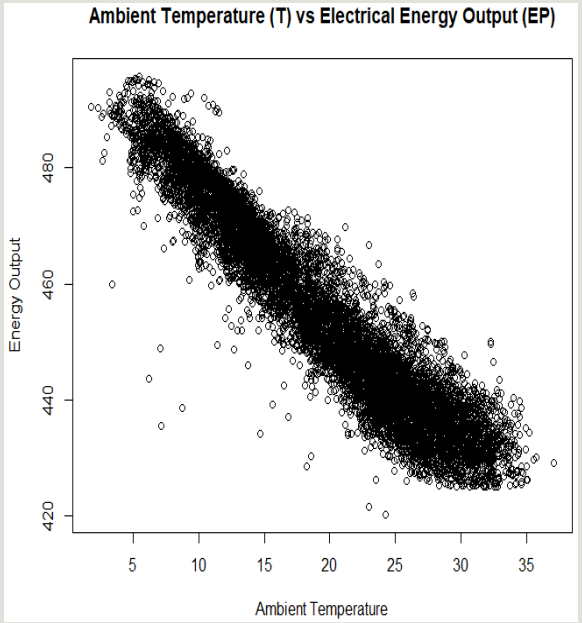


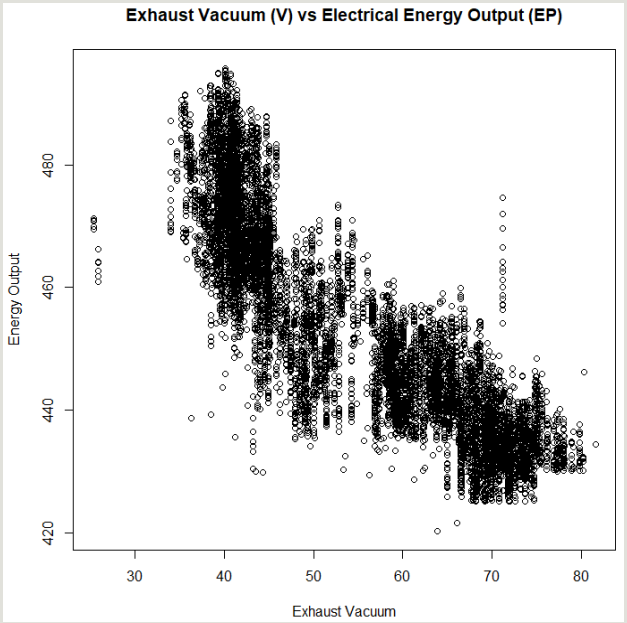
Electrical Energy Output (EP): Slight skew with lower outliers.
Ambient Temperature (AT): Evenly distributed data.
Ambient Pressure (AP): Median centered with some high outliers.
Relative Humidity (RH): Uniform distribution, no significant outliers.
Exhaust Vacuum (V): Lower quartile close to median, with some lower outliers.

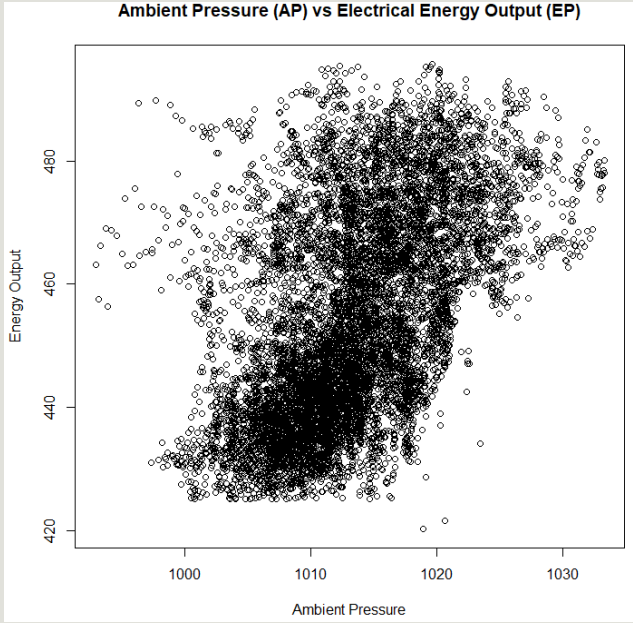
2. Data Analysis

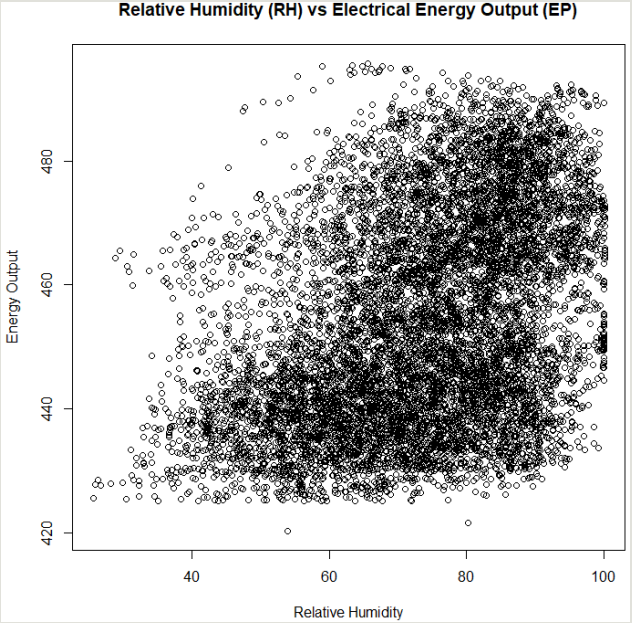
2.1 Preliminary analysis

-Scatter Plot Analysis









T&EP
Strong Negative Correlation

V&EP
Negative Correlation

AP&EP
Weak Relationship

RH&EP
Non-linear Relationship

2. Data Analysis

2.2 Principal Component Analysis (PCA) -Variance Calculation

```
> correlation_matrix <- cor(CCPP.data)
> print(correlation_matrix)
           AT          V          AP          RH          EP
AT  1.0000000  0.8441067 -0.50754934 -0.54253465 -0.9481285
V   0.8441067  1.0000000 -0.41350216 -0.31218728 -0.8697803
AP -0.5075493 -0.4135022  1.00000000  0.09957432  0.5184290
RH -0.5425347 -0.3121873  0.09957432  1.00000000  0.3897941
EP -0.9481285 -0.8697803  0.51842903  0.38979410  1.0000000
```

- EP & AT : very strong negative correlation
- EP & V : strong negative correlation
- EP & AP : weak correlation
- EP & RH : very weak correlation

Correlated Variables:

- Correlation of 1 or -1
- Convey the same information
- Including both in PCA may not be necessary or helpful
- May lead to redundant
- Consider dimensionality reduction prior to PCA

Uncorrelated Variables:

- Each contribute unique information
- Would represent distinct dimensions of the data
- Including both in PCA would be suitable

2. Data Analysis

2.2 Principal Component Analysis (PCA)

-Need for Standardization

```
> variances <- apply(CCPP.data, 2, var)
> CCPP.data.scaled <- scale(CCPP.data)
> print(variances)
       AT         V        AP        RH        EP
 55.53936 161.49054  35.26915 213.16785 291.28232
> print(head(CCPP.data.scaled))
              AT          V          AP          RH          EP
[1,] -1.51778220 -1.0651493 -0.4073356  1.14388457  1.53014579
[2,]  0.53522753  0.3292596 -0.3130402  0.06102779 -0.50477600
[3,]  1.35374774  0.2041406 -1.0286750 -2.15057533 -0.91433843
[4,] -0.07799172 -0.3632234 -1.0168880  0.23842179 -0.07470615
[5,] -1.05350680 -1.0738054  0.6518038  1.63634126  0.58973420
[6,] -0.76232829 -1.1918422  0.4699484  0.77334345  0.97234402
```

Variance in PCA:

- Measures data spread.
- Larger values indicate a wider range of data.
- Influences principal component significance.

Standardization:

- Normalizes variable scales.
- Ensures equal comparison across all variables.
- Crucial for PCA's sensitivity to variable scales.

2. Data Analysis

2.2 Principal Component Analysis (PCA)

-Performing PCA and Interpreting Results

```
> pca.result <- prcomp(CCPP.data.scaled, scale = FALSE)
> loadings <- pca.result$rotation[, 1:2]
> print(pca.result)
Standard deviations (1, .., p=5):
[1] 1.8225993 0.9563427 0.7669172 0.3737054 0.1890045

Rotation (n x k) = (5 x 5):
          PC1         PC2         PC3        PC4         PC5
AT  0.5344631  0.08033939 -0.07825384  0.3988243 0.736688717
V   0.4901834 -0.07824417 -0.45039860 -0.7420868 0.006811505
AP -0.3340928  0.59839730 -0.71253389  0.1488752 0.020838963
RH -0.2933772 -0.79026642 -0.48330099  0.1842250 0.147952945
EP -0.5257197  0.06944840  0.22300907 -0.4838840 0.659483890
> print(loadings)
          PC1         PC2
AT  0.5344631  0.08033939
V   0.4901834 -0.07824417
AP -0.3340928  0.59839730
RH -0.2933772 -0.79026642
EP -0.5257197  0.06944840
```

PCA Results

- Measures PC importance.
- Standard deviations of principal components highlight variance captured.
- Higher values => greater data variance explained.

Loadings:

- Links original variables to PCs.
- Loadings show contributions and correlations:
  - **Positive Loadings**: Positive correlation
  - **Negative Loadings**: Negative correlation.
- **PC1**: Driven by Temperature & Voltage, impacts system performance.
- **PC2**: Shows Pressure-Humidity inverse relationship, key for environmental analysis.

2. Data Analysis

2.2 Principal Component Analysis (PCA) -Percentage of Variance Explained (PVE)

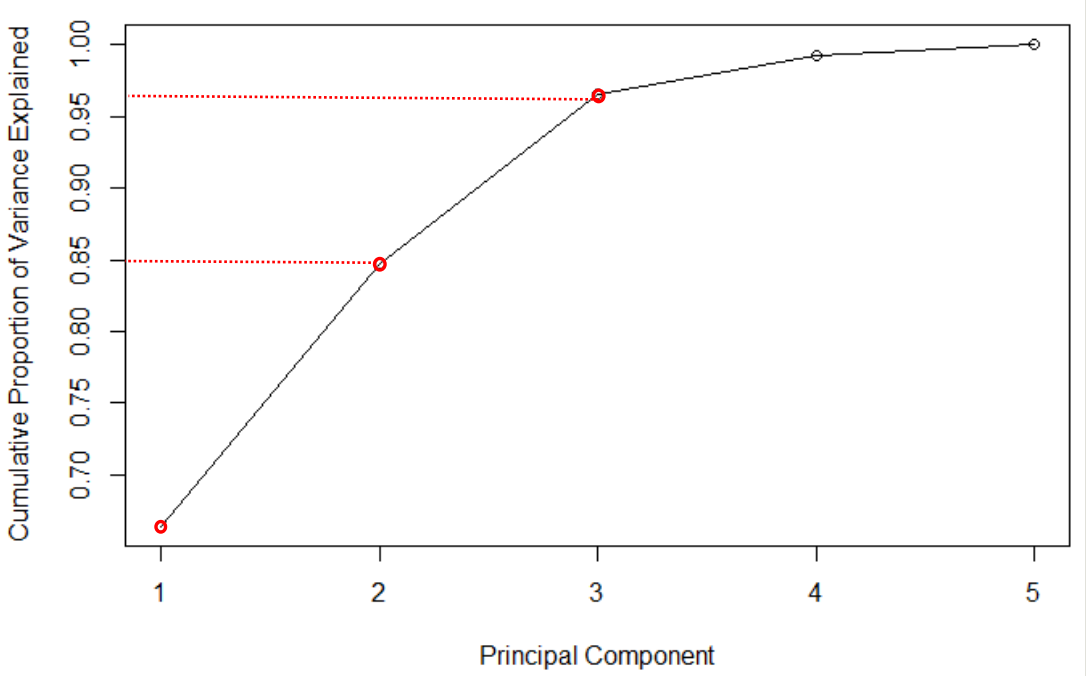
```
> pve <- pca.result$sdev^2 / sum(pca.result$sdev^2)
> cumulative.pve <- cumsum(pve)
> print(pve)
[1] 0.664373620 0.182918290 0.117632412 0.027931140 0.007144538
> print(cumulative.pve)
[1] 0.6643736 0.8472919 0.9649243 0.9928555 1.0000000
```

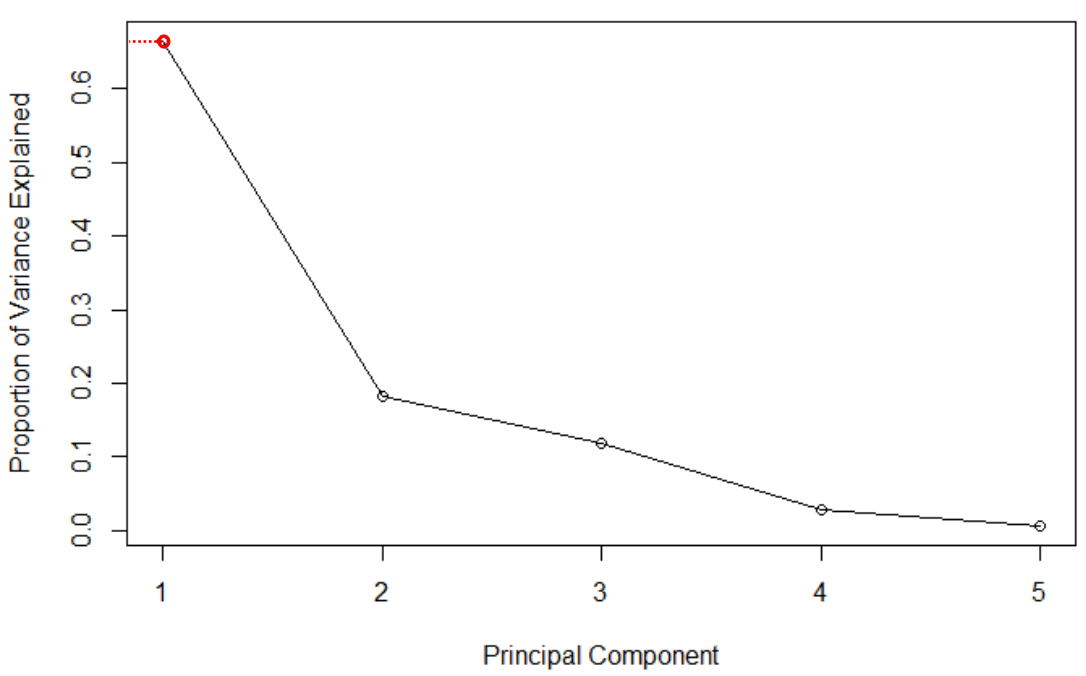
- **Primary Variance Capture**: PC1 explains approximately 66.4% of data variability.
- **Cumulative Explanation**: First two PCs account for approximately 84.7% of total variance.
- **Diminishing Returns**: Additional PCs (PC3, PC4, PC5) contribute less.
- **Dimension Reduction**: Consider using only the first two PCs for subsequent analyses based on explained variance.

2. Data Analysis

2.2 Principal Component Analysis (PCA)

-PVE Component Selection





Cumulative Proportion of Variance Explained:

The first two principal components already contain most of the information (85%).

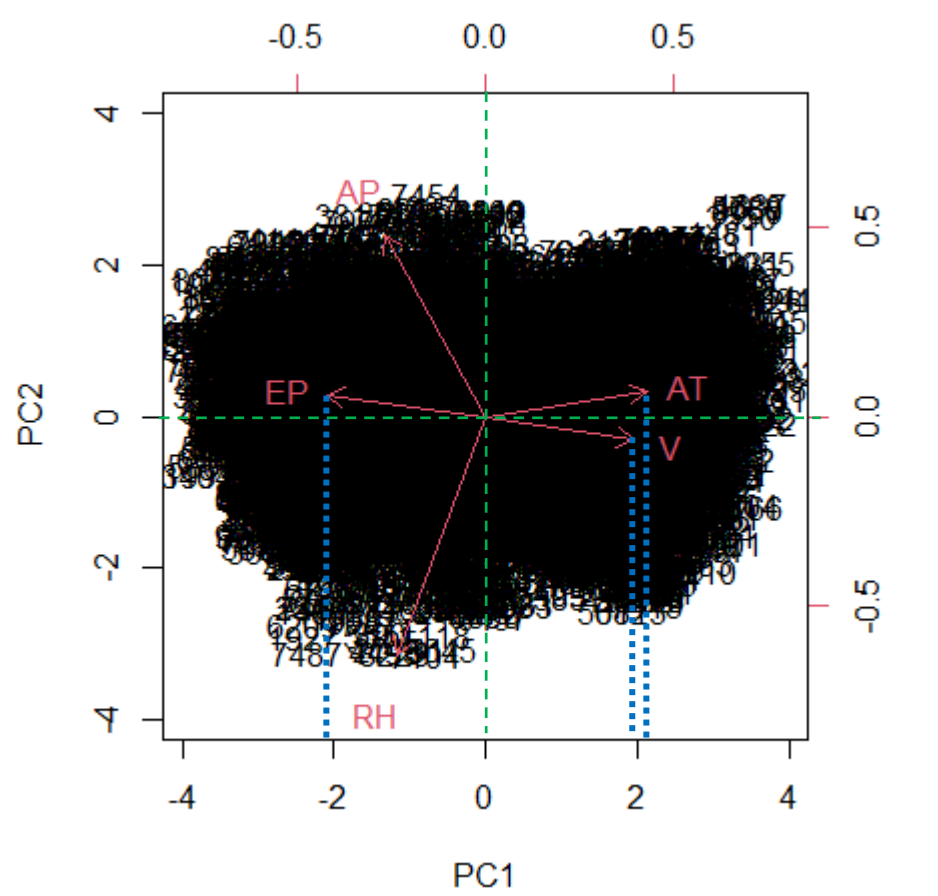
Proportion of Variance Explained:

Most of the feature information is captured by the first principal component.

2. Data Analysis

2.2 Principal Component Analysis (PCA)

-Biplot Interpretation



PCA Overview

- **Direction of Arrows:** Reflects the direction of the variable's effect in the PC space.
- **Length of Arrows:** Indicates the strength of a variable's contribution to the principal components.
- **Position of Points:** The scatter of points represents the observations.

Correlation Insights

- **Aligned**: Positively correlated.
- **Perpendicular**: Uncorrelated.
- **Opposite**: Negatively correlated.

Interpreting

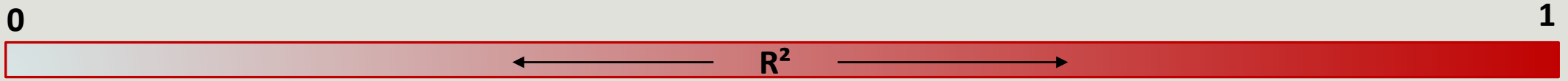
- AT & V: Positively correlated, major influencers on PC1.
- EP: Negatively associated with AT & V, affects PC1.
- AP: Positively contributes to variance on PC2.
- RH: Negatively correlated with AP, influences PC2.

2. Data Analysis

2.3 Linear Regression

Theoretical question：

In linear regression, the coefficient of determination, denoted as R^2, **measures the proportion of the variance in the dependent variable (Y)** that is predictable from the independent variables (X1 and X2).



R^2 = 0 indicates that the model explains none of the variability of the response data around its mean.

R^2 = 1 indicates that the model explains all the variability of the response data around its mean.

- R^2 is the sum of proportions of variance explained by each predictor in the model.
- However, it is not simply the arithmetic sum of the correlation coefficients (r) of the individual predictors.

2. Data Analysis

2.3 Linear Regression

	AT	V	AP	RH	EP
EP	-0.9481285	-0.8697803	0.51842903	0.38979410	1.0000000
AP	-0.5075493	-0.4135022	1.00000000	0.09957432	0.5184290
RH	-0.5425347	-0.3121873	0.09957432	1.00000000	0.3897941
V	0.8441067	1.0000000	-0.41350216	-0.31218728	-0.8697803
AT	1.0000000	0.8441067	-0.50754934	-0.54253465	-0.9481285

- The variable **most correlated with EP is the AT,** with R^2 approximately -0.948. This indicates **a strong negative correlation** between them.
- **The V** also has a **strong negative correlation** with EP, with R^2 about -0.870.
- **AP and RH** have **weaker correlations** with EP, with coefficients of 0.518 and 0.390, respectively.

2. Data Analysis

2.3 Linear Regression

```
> summary(fit)

Call:
lm(formula = EP ~ AT, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-45.951  -3.644   0.101   3.696  23.251

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 497.034120   0.156434  3177.3   <2e-16 ***
AT           -2.171320   0.007443  -291.7   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.426 on 9566 degrees of freedom
Multiple R-squared:  0.8989,    Adjusted R-squared:  0.8989
F-statistic: 8.51e+04 on 1 and 9566 DF,  p-value: < 2.2e-16
```

- **The estimated intercept coefficient** is 497.034.

➡ This represents the expected value of the EP when the AT is zero.

- **The estimated coefficient** (βˆ1)for ambient temperature is -2.171.

➡ This means that for every one unit increase AT, EP is expected to decrease by 2.171 units.

2. Data Analysis

2.3 Linear Regression

$$\hat{\beta}_1 \pm t_{\alpha/2,df} \times SE(\hat{\beta}_1)$$

- **the general expression of a 1 – α confidence**

```
> confint(fit, level = 0.95)
                 2.5 %    97.5 %
(Intercept) 496.72748 497.34076
AT           -2.18591  -2.15673
```

- **95% confidence level is fall between -2.1859 and -2.1567.**
  ➡️ This interval does not contain zeros, indicating that the relationship between AT and EP is statistically significant.

2. Data Analysis

2.3 Linear Regression

```
Call:
lm(formula = EP ~ AT, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-45.951  -3.644   0.101   3.696  23.251

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 497.034120   0.156434  3177.3   <2e-16 ***
AT           -2.171320   0.007443  -291.7   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.426 on 9566 degrees of freedom
Multiple R-squared:  0.8989,    Adjusted R-squared:  0.8989
F-statistic: 8.51e+04 on 1 and 9566 DF,  p-value: < 2.2e-16
```

- The estimated β1 : -2.171.
  The standard error: 0.007.
  T-value: -291.7.
  p-value is less than 2e-16.

➡ Very low p-value means that we can **reject the null hypothesis** at a very high level of confidence

➡ β^1 is significantly non-zero, which indicates **AT has a significant effect EP**

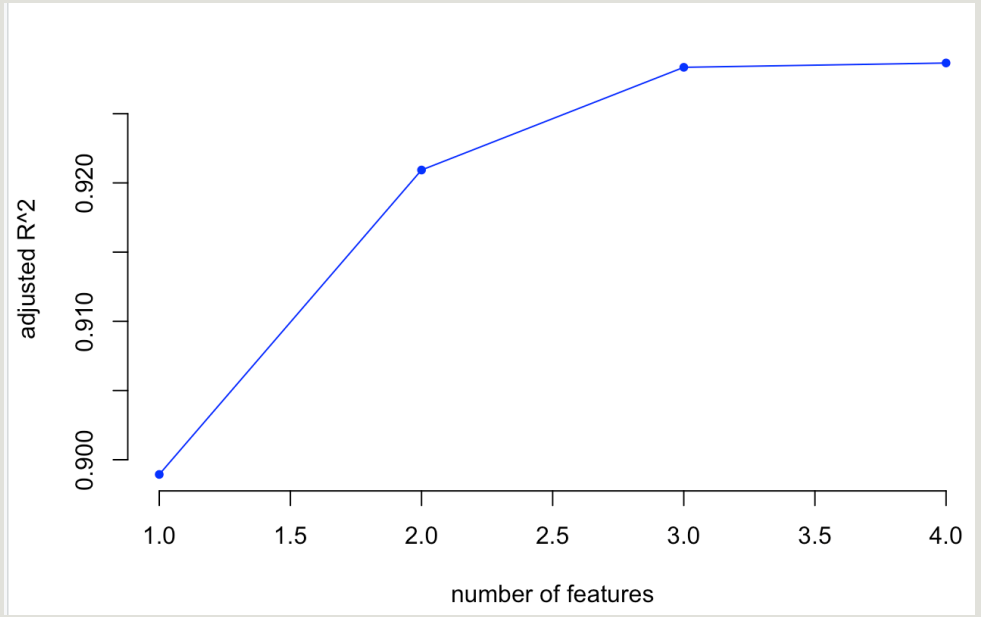
- **The value of the coefficient of determination R^2** is 0.8989

➡ the model explains approximately 89.89% of the variability in EP.

2. Data Analysis

2.3 Linear Regression

```
> summary(regfit.full)
Subset selection object
Call: regsubsets.formula(EP ~ ., data = data, nvmax = 4)
4 Variables  (and intercept)
   Forced in Forced out
AT     FALSE      FALSE
V      FALSE      FALSE
AP     FALSE      FALSE
RH     FALSE      FALSE
1 subsets of each size up to 4
Selection Algorithm: exhaustive
         AT  V   AP  RH
1  ( 1 ) "*" " " " " " "
2  ( 1 ) "*" " " " " "*"
3  ( 1 ) "*" "*" " " "*"
4  ( 1 ) "*" "*" "*" "*"
```



1) Best mode: four features , 0.929
2) four features： AT, V, AP, RH

2. Data Analysis

2.3 Linear Regression

- **Disadvantages of R square**: The value of R^2 is always between 0 and 1 and and does not decrease as more predictor variables are added to the model.

- **Advantages of adjusted R square**: adjusting for the number of predictors in the model. It compensates for the addition of variables and only increases if the new variable improves the model more than would be expected by chance.

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

- **To conclude**, it is more appropriate to use adjusted R^2 when comparing models that contain different numbers of predictor variables

2. Data Analysis

2.3 Linear Regression

```
> summary(model)

Call:
lm(formula = EP ~ AT + V + AP + RH, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-43.435  -3.166  -0.118   3.201  17.778

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) 454.609274   9.748512   46.634  < 2e-16 ***
AT           -1.977513   0.015289 -129.342  < 2e-16 ***
V            -0.233916   0.007282  -32.122  < 2e-16 ***
AP            0.062083   0.009458    6.564 5.51e-11 ***
RH           -0.158054   0.004168  -37.918  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.558 on 9563 degrees of freedom
Multiple R-squared:  0.9287,    Adjusted R-squared:  0.9287
F-statistic: 3.114e+04 on 4 and 9563 DF,  p-value: < 2.2e-16
```

- Intercept: 454.609

➡ This represents the expected energy output (EP) value of 454.609 when AT, V, AP, and RH are all zero.

- AT: -1.978, negative
- V: -0.234, negative
- AP: 0.062 positive
- RH: -0.158, negative

- R^2: 0.929

➡ 0.929 is a **very high value** indicating that the model fits the data very well and shows a **strong correlation** between the variables.

2. Data Analysis

2.3 Linear Regression

```
Call:
lm(formula = EP ~ AT + V + AP + RH, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-43.435  -3.166  -0.118   3.201  17.778

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) 454.609274   9.748512   46.634  < 2e-16 ***
AT           -1.977513   0.015289 -129.342  < 2e-16 ***
V            -0.233916   0.007282  -32.122  < 2e-16 ***
AP            0.062083   0.009458    6.564 5.51e-11 ***
RH           -0.158054   0.004168  -37.918  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.558 on 9563 degrees of freedom
Multiple R-squared:  0.9287,    Adjusted R-squared:  0.9287
F-statistic: 3.114e+04 on 4 and 9563 DF,  p-value: < 2.2e-16
```

β1 for AT is -1.978, t-value is -129.342, p-value is < 2e-16.
β2 for V is -0.234, t-value is -32.122, p-value is < 2e-16.
β3 for AP is 0.062, t-value is 6.564, p-value is 5.51e-11.
β4 for RH is -0.158, t-value is -37.918, p-value is < 2e-16.

- Conclusion: The coefficients of each of the predictor variables are significantly different from zero and all have a significant effect on predicted energy output

2. Data Analysis

2.3 Linear Regression

```
> new_data <- data.frame(AT = 22, V = 75, AP = 1010, RH = 80)
> predicted_EP <- predict(model, new_data)
> predicted_EP
       1
443.6197
```

When AT = 22°C, V = 75, AP = 1010, RH = 80%, the predicted net hourly electrical power output (EP) would be approximately 443.62 MW.

3. Conclusion

Preliminary analysis

Principal Component Analysis (PCA)

Simple Linear Regression

Feature selection for multiple linear regression

4. References

— Tu¨fekci,Pinar and Kaya,Heysem. (2014). Combined Cycle Power Plant. UCI Machine Learning Repository. https://doi.org/10.24432/C5002N.

— Tu¨fekci,Pinar (2014) : ”Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods”, International Journal of Electrical Power & Energy Systems, Volume 60, September 2014, pp 126-140, ISSN 0142-0615, http: //www.sciencedirect.com/science/article/pii/S0142061514000908.

— Kaya Heysem, Tu¨fekci Pinar and Fikret Gu¨rgen Sadik (2012) : ”Local and Global Learning Methods for Predicting Power of a Combined Gas & Steam Turbine”, Proceedings of the International Conference on Emerging Trends in Computer and Electronics Engineering ICETCEE 2012, pp. 13-18,Dubai