Symbiosis Centre of Management Studies

NOIDA

Assessment Component-2

(2020-21)

Predictive Analytics

Aman Tibrewal (18021021023)

Divye Kant (18021021089)

Vridhi Wadhawan (18021021354)

Submitted to:

Dr Kriti Priya Gupta

Associate Professor

Objective: To formulate an accurate predictive linear relation model for the Average Monthly Working Hours in the Employee Data File

Case A: Employee hours

Company ABC has very poor employee satisfaction and retention. They have recently conducted a series of exit interviews to understand what went wrong and how they could make an impact on employee retention. A large number of the leaving employees indicated that they would have stayed if they were compensated with overtime pay for their extra hours. While Company ABC may not have been tracking employee hours this year, they do have a sample of previous employee data from an in-depth employee quiz performed 2 years ago. The information available for the sample employees includes currently available information such as:

S.No.	Variable	Description	Data Type
1.	satisfaction_level	Satisfaction level of employees	Continuous
2.	last_evaluation	Evaluation received in the last appraisal	Continuous
3.	number_project	Number of projects worked on	Discrete
4.	average_montly_hours	Average hours an employee is likely to work	Continuous
5.	time_spend_company	Tenure (in years)	Continuous
6.	Work_accident	Whether accident happened during work (Yes=1, No=0)	Nominal
7.	left	Left the company or not (Yes=1, No=0)	Nominal
8.	promotion_last_5years	Whether received promotion during last 5 years (Yes=1, No=0)	Nominal
9.	dept	Department where employee is working	Nominal
10.	salary	Salary level	Nominal

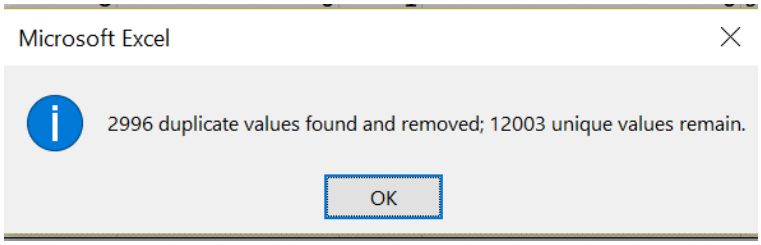
The company wants to estimate the average hours an employee is likely to work so that it can estimate how much money it would have to pay out to its employees. Specifically, the company wants to build a predictive model for the average hours an employee is likely to work based on the other factors.

The related to the analysis and the working directory can be found in:

https://github.com/Aman-Tibrewal/Predictive-Component

Data Cleaning and Processing:

There were 2996 duplicate entries in the dataset, which were removed from the dataset, after which the dataset was left with 12003 rows of data.



In addition to this, there were certain missing values identified, which were treated accordingly.

Variable	Data Type	Mean	Mode	Missing Values	Method of Treating Missing Values
Satisfaction Level	Continuous	0.63	0.74	3.00	Imputation (Mean Value)
Last Evaluation	Continuous	0.72	0.55	4.00	Imputation (Mean Value)
Number Project	Discrete	3.80	4.00	0.00	Imputation (Mean Value)
Average Monthly Hours	Continuous	200.47	160.00	3.00	Imputation (Mean Value)
Time Spend Company	Continuous	3.37	3.00	0.00	Imputation (Mean Value)
Work Accident	Nominal	0.15	0.00	1.00	Imputation (Mode Value)
Left	Nominal	0.17	0.00	3.00	Imputation (Mode Value)
Promotion Last 5 Years	Nominal	0.02	0.00	0.00	Imputation (Mode Value)
Dept	Nominal	6.97	8.00	1.00	Imputation (Mode Value)
Salary	Nominal	1.60	1.00	0.00	Imputation (Mode Value)

Assumptions Testing:

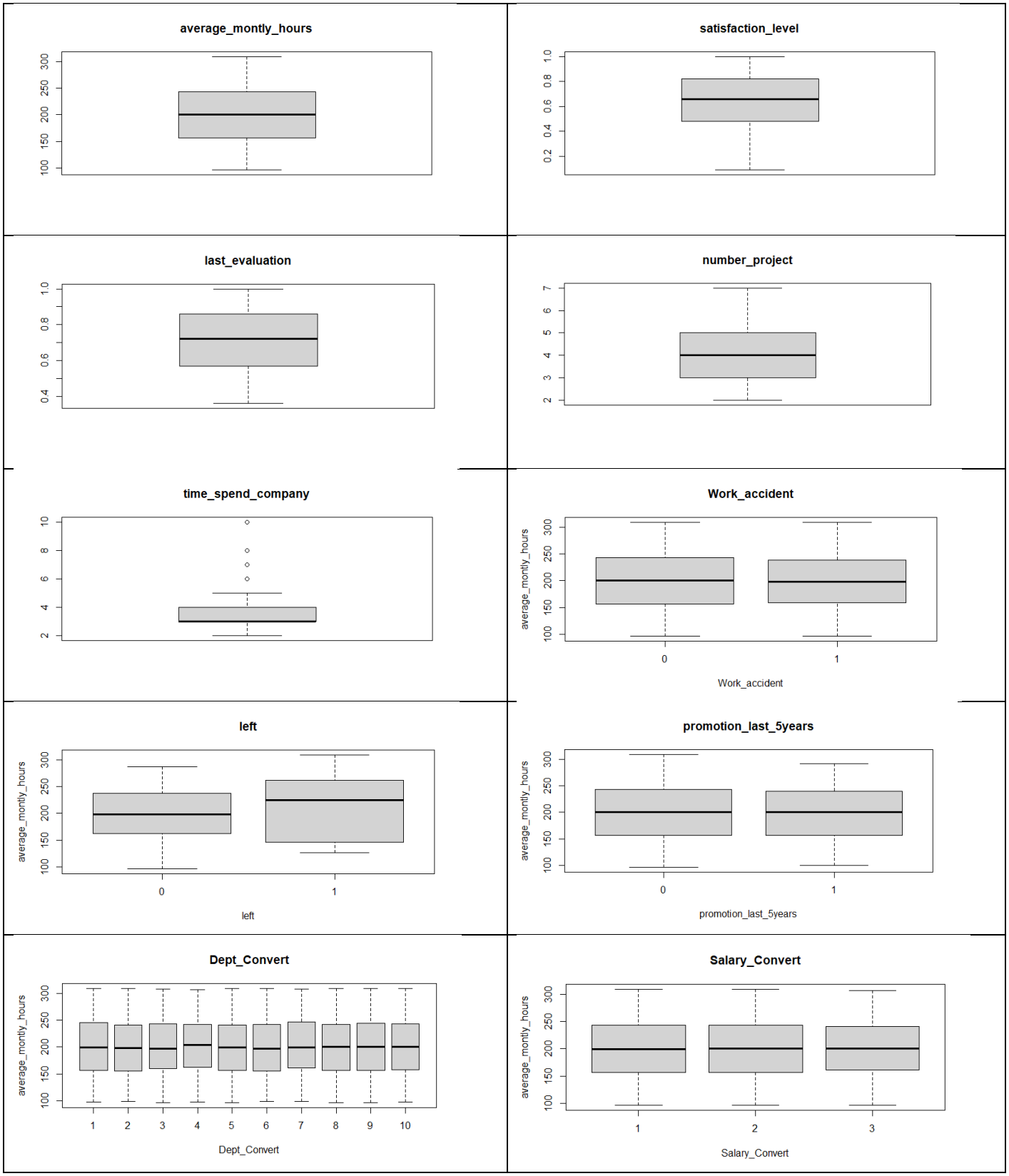
Normality Testing

Variable	Kurtosis	Skewness
satisfaction_level	2.477300	-0.539013
last_evaluation	1.820919	-0.031801
number_project	2.604474	0.332810
average_montly_hours	1.916970	0.027523
time_spend_company	8.132913	1.815038
Log_TimeSpent	3.185936	0.544127
Inv_TimeSpent	2.191859	0.236493
Imp_timeSpend	2.841937	0.613427

For a variable to be normally distributed, the skewness and kurtosis values have to be within the following threshold: Skewness between -2 and +2 and Kurtosis between -7 and +7 for R.

Thus, as per the Descriptive Statistics reported for the variables in Table 1, all the variables expect the Time Spent Company Variable are normally distributed. The Time Spent variable has extremely high skewness and kurtosis.

Outliers Test



As per Table _, the boxplots for all the variables, it can be established that most of the data does not have many outliers, except Time Spend Company.

Processing Heavy Skewness and Outliers

Since the data for Time Spend in the company is heavily skewed and has 824 out of 12003 rows as outliers, we have transformed it using the following transformations:

Variable	Transformation	Outliers Treated	Kurtosis	Skewness
Log_TimeSpent	Log10	False	3.185936	0.544127
Inv_TimeSpent	Inverse	False	2.191859	0.236493
Imp_TimeSpend	Imputation (With Mean Values)	True	2.841937	0.613427

Linear Relationship and Multicollinearity Test

Final Selection	Average Monthly hours	Satisfaction level	Last evaluation	Number project	Time spend company	Log Time Spent	Imp time Spend	Inv Time Spent
Average Monthly hours	1	-0.006	0.265**	0.331**	0.103**	0.106**	0.139**	-0.094**
Satisfaction level	-0.006	1	0.095**	-0.134**	-0.153**	-0.187**	-0.211**	0.202**
Last evaluation	0.265**	0.095**	1	0.271**	0.097**	0.096**	0.132**	-0.081**
Number project	0.331**	-0.134**	0.271**	1	0.189**	0.202**	0.248**	-0.189**
Time spend company	0.103**	-0.153**	0.097**	0.189**	1	0.967**	0.63**	-0.893**
Log Time Spent	0.106**	-0.187**	0.096**	0.202**	0.967**	1	0.768**	-0.977**
Imp time Spend	0.139**	-0.211**	0.132**	0.248**	0.63**	0.768**	1	-0.83**
Inv Time Spent	-0.094**	0.202**	-0.081**	-0.189**	-0.893**	-0.977**	-0.83**	1

As per Table _, It can be established that all the data is significantly correlated with *Average Monthly Hours*, except the variables (Marked with *Red* in the table): *Satisfaction Level*; This is because the p-value for the variable was greater than 0.05 (i.e., at 95% confidence interval). Thus, it can be established that the *Average Monthly Hours* Variable has a linear relationship with the *3* other variables (i.e., Last Evaluation, Number Project, Time Spend) as well as the conversions for Time Spent (Log, Inverse and Imputed).

Additionally, as it can be noticed in Table _, it can be said that there is a situation of multicollinearity between 3 variables (Marked in Yellow in the table). But that is because of the variables are the transformed versions of each other, thus, resulting in multicollinearity amongst each other.

Testing the Linear Relation between the Categorical Variables and the Dependant Variable

Department

Levene's test (Levene statistic=0.8101, p=0.6068>0.05) indicates that the variances of Average Monthly Hours in the 10 groups of Department are homogeneous (no significant difference in variance). Hence, Welch ANOVA will not be used, simple ANOVA will be used. As per the ANOVA Results (F= 0.175 and p=0.676>0.05), Since the p-value is more than 0.05 then we can conclude that there is no linear relation between the department and the average monthly hours as there is no significant impact of the 2 on each other.

Salary

Levene's test (Levene statistic= 3.557, p= 0.02856<0.05) indicates that the variances of Average Monthly Hours in the 3 groups of Salary are not homogeneous (significant difference in variance). Hence, Welch ANOVA will be used, simple ANOVA will not be used. As per the Welch ANOVA Results (F= 2804.0 and p=0.951>0.05), Since the p-value is more than 0.05 then we can conclude that there is no linear relation between the salary and the average monthly hours as there is no significant impact of the 2 on each other.

Work Accident

Levene's test (Levene statistic= 13.89, p=0.00019<0.05) indicates that the variances of prices in the 2 groups of Work Accident are not homogeneous (significant difference in variance). Hence, it is established that there are no equal variances. As per the Welch 2 Sample T-Test Results (T= 1.4445 and p=0.1487 >0.05), Since the p-value is more than 0.05 then we can conclude that there is no linear relation between average monthly hours and Work Accidents as there is no significant impact of the 2 on each other.

Left

Levene's test (Levene statistic= 1085.8, p= 2.2e-16<0.05) indicates that the variances of prices in the 2 groups of Left are not homogeneous (significant difference in variance). Hence, it is established that there are no equal variances. As per the Welch 2 Sample T-Test Results (T= -6.2841 and p= 3.887e-10 <0.05), Since the p-value is less than 0.05 then we can conclude that there is a linear relation between average monthly hours and Left as there is a significant impact of the 2 on each other.

Promotion last 5 years

Levene's test (Levene statistic= 0.2031, p= 0.6522 >0.05) indicates that the variances of prices in the 2 groups of Promotion last 5 years are homogeneous (no significant difference in variance). Hence, it is established that there are equal variances. As per the Welch 2 Sample T-Test Results (p=0.5882(>0.05) and T=0.54143), Since the p-value is more than 0.05 then we can conclude that there is no linear relation between average monthly hours and Work Accidents as there is no significant impact of the 2 on each other.

Final Acceptable Variables Results as per the Assumptions:

Final Selection	Normality	Outliers	Linear Relationship	Multicollinearity
satisfaction_level	TRUE	TRUE	FALSE	TRUE
last_evaluation	TRUE	TRUE	TRUE	TRUE
number_project	TRUE	TRUE	TRUE	TRUE
time_spend_company	TRUE	FALSE	TRUE	FALSE (1)
Work_accident	TRUE	TRUE	FALSE	TRUE
left	TRUE	TRUE	TRUE	TRUE
promotion_last_5years	TRUE	TRUE	FALSE	TRUE
Dept_Convert	TRUE	TRUE	FALSE	TRUE
Salary_Convert	TRUE	TRUE	FALSE	TRUE
Log_TimeSpent	TRUE	FALSE	TRUE	FALSE (1)
Inv_TimeSpent	TRUE	FALSE	TRUE	FALSE (1)
Imp_TimeSpent	TRUE	TRUE	TRUE	TRUE

Therefore, the final selected variables for independent variables are:

- Last Evaluation
- Number of Projects
- Left
- Time Spent (Log, Inverse and Imputed)

Model Testing:

MODEL 1

	F (4,11998)	Prob > P	R-Squared	Adj-R2	
	516.492	2.2e-16	0.147	0.1469	
	Estimate	**Std. Error**	**t value**	**PR(>|t|)**	**Std. Beta Coeff.**
(Intercept)	112.082	2.29	48.948	2E-16	
last_evaluation	53.998	2.542	21.24	2E-16	0.18654257
number_project	11.349	0.376	30.158	2E-16	0.27123093
left	6.297	1.164	5.409	6.5E-08	0.04816194
Imp_timeSpend	1.761	0.517	3.41	0.00065	0.03139965

Variables Taken:

DV: Average_monthly_hours;

IV: Last_evaluation, Number_project, Left, Imp_TimeSpent

Model significance: F=516.492, $p<0.05$ indicate that overall regression model is significant

Significance of individual predictors:

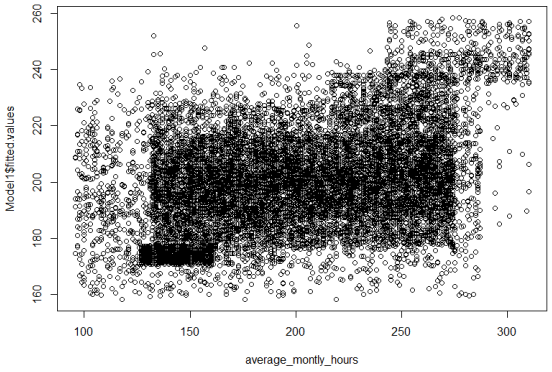
- The Table indicates that all the IV's have a significant effect on the DV. Hence Average_monthly_hours are dependent on all IV's
- last_evaluation and number_project are the most significant predictors of Average_monthly_hours followed by left and Imp_timeSpend
- All IVs are affecting the DV positively.

Model Fit:

Adjusted R^2= 14.7%

Standard Error = 45.01 on 11998 degrees of freedom

The model is able to explain 14.7% variation in Average_monthly_hours.



MODEL 2

	F (3,11999)	Prob > P	R-Squared	Adj-R2	
	684.174	2.2e-16	0.1461	0.1459	
	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	**Std. Beta Coeff.**
(Intercept)	115.79	2.016	57.436	2E-16	
last_evaluation	54.623	2.537	21.532	2E-16	0.18870057
number_project	11.638	0.367	31.722	2E-16	0.27812347
Left	7.565	1.104	6.855	4.9E-09	0.05786218

Variables Taken:

DV: Average_monthly_hours;

IV: Last_evaluation, Number_project, Left

Model significance: F= 684.2, p<0.05 indicate that overall regression model is significant

Significance of individual predictors:

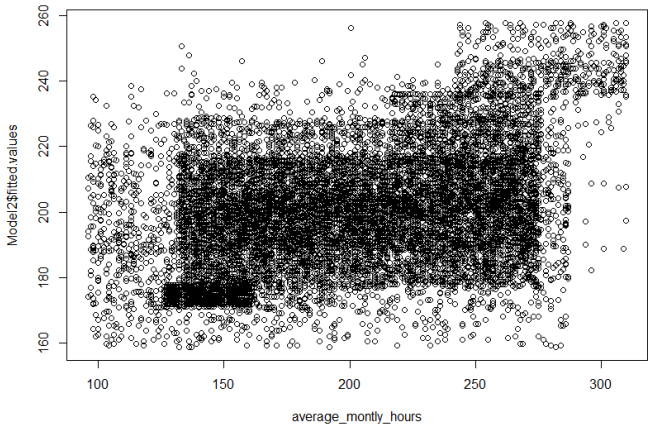
- The Table indicates that all the IV's have a significant effect on the DV. Hence Average_monthly_hours are dependent on all IV's
- last_evaluation and number_project are the most significant predictors of Average_monthly_hours followed by left.
- All IVs are affecting the DV positively.

Model fit

Adjusted R^2 =14.59%

Standard Error = 45.03 on 11999 degrees of freedom

Hence, Model 2 is worse off than Model 1 as it is able to explain a lesser degree of the total variation in DV.



MODEL 3

	F (4,11998)	Prob > P	R-Squared	Adj-R2	
	513.69	2.2e-16	0.1462	0.1459	
	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	**Std. Beta Coeff.**
(Intercept)	118.208	2.628	44.977	2E-16	
last_evaluation	54.508	2.538	21.477	2E-16	0.18830178
number_project	11.546	0.372	31.005	2E-16	0.27592955
Left	7.152	1.141	6.271	3.71E-10	0.05470242
Inv_TimeSpent	-5.715	3.986	-1.434	0.152	-0.01273642

Variables Taken:

DV: Average_monthly_hours;

IV: Last_evaluation, Number_project, Left, Inv_TimeSpent

Model significance: F= 513.7, p<0.05 indicate that overall regression model is significant

Significance of individual predictors:

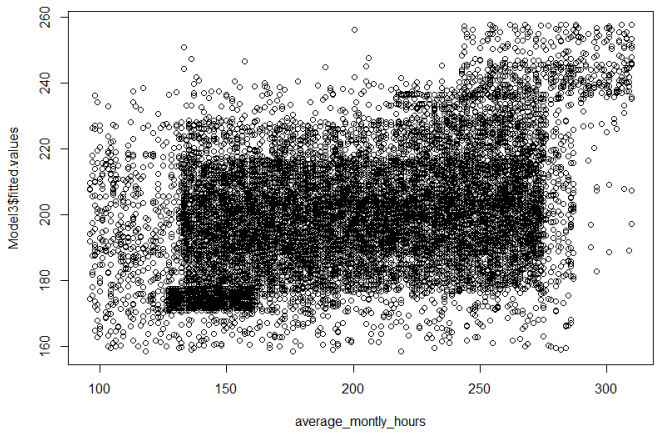
- The Table indicates that all the IV's have a significant effect on the DV except for Inv_TimeSpent. Hence Average_monthly_hours are dependent on all IV's except for Inv_TimeSpent
- last_evaluation and number_project are the most significant predictors of Average_monthly_hours followed by left.
- All IVs are influencing the DV positively except for Inv_TimeSpent which is negatively influencing it.

Model fit

Adjusted R^2 = 14.59%

Standard Error = 45.03 on 11998 degrees of freedom

Hence, Model 3 is still worse off than the original Model 1 in being able to explain the degree of variation in the DV. But its results are equivalent to Model 2.



Redacting Outlier Rows from Employee Data

After Model Testing with the Original dataset, there was another Dataset created, after removing the 824 rows of data from the 12003 rows, which were the outliers in the Time Spent in the Company Variable. The resulting dataset was named as EmployeeData_Redacted and in the variable names, "2" was concatenated in the end of all the variable names.

Additionally, since there were various rows of data which were removed, the assumptions testing was done again in the new data frame to check if there were any new developments in the assumptions.

The results were as followed:

Assumption Testing (Redacted):

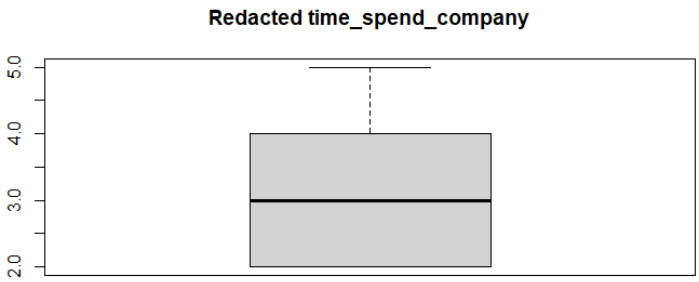
Normality Test Redacted	Skewness	Kurtosis
satisfaction_level2	-0.5514	2.5383
last_evaluation2	-0.0154	1.8175
number_project2	0.3559	2.6467
average_montly_hours2	0.0571	1.9156
time_spend_company2	0.5678	2.6328
Log_TimeSpent2	0.0383	2.1677
Inv_TimeSpent2	0.4081	2.0244

As per the above table, all the variables were normally distributed along with the New Time Spend and their transformations.

Correlation Test Redacted	Average Monthly hours 2	Satisfaction level 2	Last evaluation 2	Number project 2	Time Spend company 2	Log Time Spent 2	Inv Time Spent 2
Average Monthly hours 2	1.000	-0.012	0.267	0.344	0.145	0.123	0.222
Satisfaction level 2	-0.012	1.000	0.095	-0.134	-0.222	-0.226	-0.087
Last evaluation 2	0.267	0.095	1.000	0.281	0.138	0.114	-0.195
Number project 2	0.344	-0.134	0.281	1.000	0.259	0.230	-0.099
Time Spend company 2	0.145	-0.222	0.138	0.259	1.000	0.988	-0.954
Log Time Spent 2	0.123	-0.226	0.114	0.230	0.988	1.000	-0.989
Inv Time Spent 2	0.222	-0.087	-0.195	-0.099	-0.954	-0.989	1.000

As per the above table, it can be established that there are no new major developments in the correlations among the redacted data. Just like the original file, it can be witnessed that there isn't a linear relation b/w the

Dependant Variable (Average Monthly Hours) and Satisfaction Level, and there is a linear relation between the various types of Time Spend Data and their Transformations.



As per the above boxplot, it can be established that all the outliers from the original Time Spend Variable have been removed.

Linear Relation for Categorical Variations				
Variable	**Test**	**Test Value**	**P Value**	**Interpretation**
Work Accident 2	Levene's Test	14.369	0.000151	No Homogeneity
	Welch T Test	0.808	0.419	No Significant Impact
Left 2	Levene's Test	1230.100	2.20E-16	No Homogeneity
	Welch T Test	-4.758	2.08E-06	Significant Impact
Promotion Last 5 Years	Levene's Test	0.320	0.5717	Homogeneity
	Independent T Test	0.924	0.3556	No Significant Impact
Department 2	Levene's Test	0.863	0.5583	Homogeneity
	ANOVA	0.137	0.711	No Significant Impact
Salary 2	Levene's Test	4.255	0.01421	No Homogeneity
	Welch ANOVA	2505.000	0.912	No Significant Impact

Since all the assumptions are similar to the original dataset, thus the selected variables are as followed:

- Last Evaluation
- Number of Projects
- Left
- Time Spent (Original, Log and Inverse)

MODEL 4

	F (4,11174)	Prob > P	R-Squared	Adj-R2	
	507.462	2.2e-16	0.1537	0.1534	
	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	**Std. Beta Coeff.**
(Intercept)	110.873	2.317	47.86	2E-16	
last_evaluation2	53.226	2.637	20.183	2E-16	0.18367906
number_project2	11.804	0.389	30.353	2E-16	0.28330932
left2	5.224	1.196	4.367	1.3E-05	0.04033066
time_spend_company2	1.792	0.517	3.465	0.00053	0.03319568

Variables Taken:

DV: Average_monthly_hours;

IV: Last_evaluation, Number_project, Left, Time Spent

Model significance: F=507.462, p<0.05 indicate that overall regression model is significant

Significance of individual predictors:

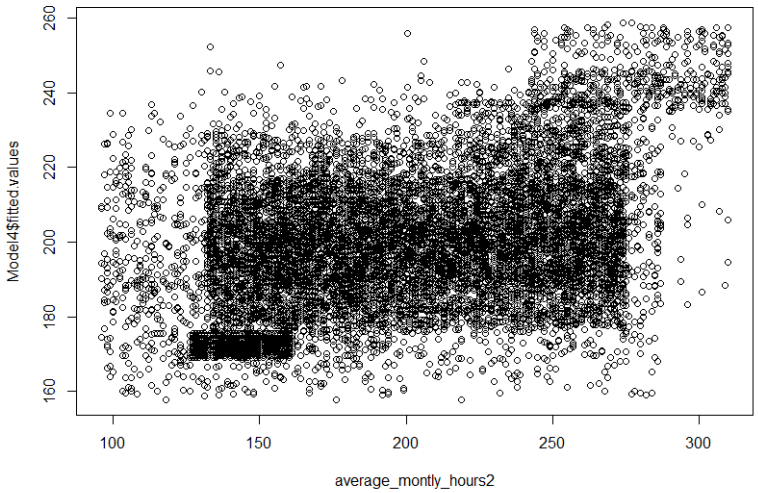
- The Table indicates that all the IV's have a significant effect on the DV. Hence Price Average_monthly_hours are dependent on all the variables.
- last_evaluation2 and number_project2 are the most significant predictors of Average_monthly_hours followed by left2 and time_spend_company 2.
- All the IV's positively influence the DV

Model Fit:

Adjusted R^2 =15.34%

Standard Error = 44.68 on 11174 degrees of freedom

The model is able to explain 15.34% variation in Average_monthly_hours.



MODEL 5

	F (3,11175)	Prob > P	R-Squared	Adj-R2	
	671.953	2.2e-16	0.1528	0.1526	
	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	**Std. Beta Coeff.**
(Intercept)	114.472	2.072	55.254	2E-16	
last_evaluation2	53.979	2.63	20.528	2E-16	0.18627508
number_project2	12.121	0.378	32.054	2E-16	0.29092314
left2	6.606	1.128	5.855	4.90E-09	0.05100475

Variables Taken:

DV: Average_monthly_hours;

IV: Last_evaluation, Number_project, Left

Significance of individual predictors:

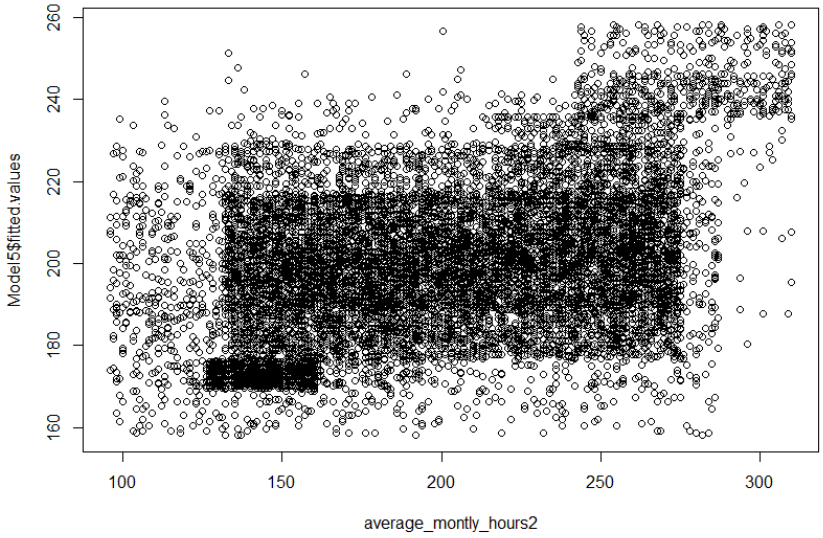
- The Table indicates that all the IV's have a significant effect on the DV. Thus, Average_monthly_hours is dependent on all IV's.
- left2 is the most significant predictors of Average_monthly_hours followed by last_evaluation2 and number_project2.
- All IVs are affecting the DV positively.

Model fit

Adjusted R^2 =15.26%

Standard Error = 44.7 on 11175 degrees of freedom

Hence, Model 5 is slightly worse off than Model 4 as it is able to explain a lesser degree of the total variation in the DV



MODEL 6

	F (4,11174)	Prob > P	R-Squared	Adj-R2	
	505.561	2.2e-16	0.1532	0.1529	
	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	**Std. Beta Coeff.**
(Intercept)	111.5	2.424	45.989	2.00E-16	
last_evaluation2	53.589	2.634	20.344	2.00E-16	0.18493226
number_project2	11.93	0.387	30.85	2.00E-16	0.28632424
left2	5.684	1.194	4.761	1.95E-06	0.0438863
Log_TimeSpent2	8.701	3.69	2.358	0.0184	0.02235662

Variables Taken:

DV: Average_monthly_hours;

IV: Last_evaluation, Number_project, Left, Log_TimeSpent

Model fit

Adjusted R^2 =15.29%

Standard Error = 44.69 on 11174 degrees of freedom

Significance of individual predictors:

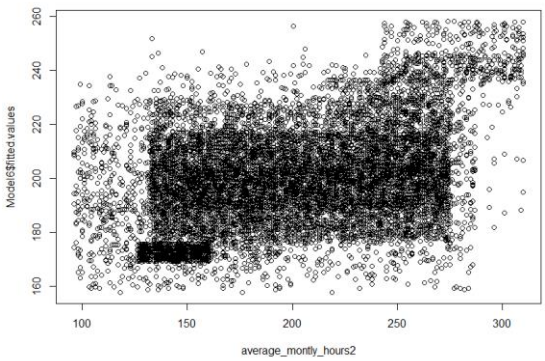
- The Table indicates that all the IV's have a significant effect on the DV. Thus, Average_monthly_hours is dependent on all IV's.
- last_evaluation2 and number_project2 are the most significant predictors of Average_monthly_hours followed by left2 and Log_TimeSpent2.
- All IVs are affecting the DV positively.

Model fit

Adjusted R^2 =15.29%

Standard Error = 44.69 on 11174 degrees of freedom

Hence, Model 6 is slightly better than Model 5 as it is able to explain a lesser degree of the total variation in the DV but is inferior to model 4 in it.



MODEL 7

F (4,11174)	Prob > P	R-Squared	Adj-R2
504.337	2.2e-16	0.1529	0.1526

	Estimate	Std. Error	t value	Pr(>|t|)	Std. Beta Coeff.
(Intercept)	116.848	2.877	40.611	2.00E-16	
last_evaluation2	53.844	2.632	20.458	2.00E-16	0.18581242
number_project2	12.039	0.384	31.325	2.00E-16	0.28895972
left2	6.16	1.189	5.181	2.25E-07	0.04755837
Inv_TimeSpent2	-5.428	4.561	-1.19	0.234	-0.01113823

Variables Taken:

DV: Average_monthly_hours;

IV: Last_evaluation, Number_project, Left, Inv_TimeSpent

Significance of individual predictors:

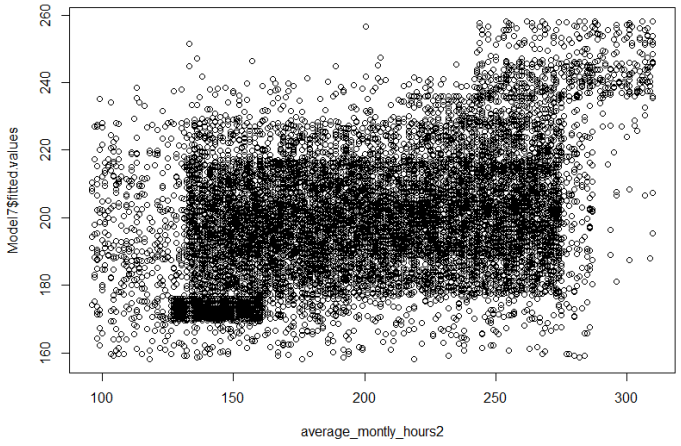
- The Table indicates that all the IV's have a significant effect on the DV. Thus, Average_monthly_hours is dependent on all IV's.
- last_evaluation2, number_project2 and left2 are the most significant predictors of Average_monthly_hours followed by Inv_TimeSpent2.
- last_evaluation2, number_project2 and left2 are affecting the DV positively where as Inv_TimeSpent2 is affecting the DV negatively.

Model fit

Adjusted R^2 =15.26%

Standard Error = 44.7 on 11174 degrees of freedom

Hence, Model 7 is slightly worse off than Model 6 as it is able to explain a lesser degree of the total variation in the DV. It is equivalent to model 5 and inferior to model 4.



Summary

Models	F Value	P Value	Adj R2
Model 1	516.492	2.20E-16	0.147
Model 2	684.174	2.20E-16	0.1459
Model 3	513.69	2.20E-16	0.1459
Model 4	507.462	2.20E-16	0.1534
Model 5	671.953	2.20E-16	0.1526
Model 6	505.561	2.20E-16	0.1529
Model 7	504.337	2.20E-16	0.1526

Final Selected Model: MODEL 4

DV: Average_monthly_hours;

IV: Last_evaluation, Number_project, Left, Time Spent

Model Fit:

Adjusted R^2 =15.34%

Standard Error = 44.68 on 11,174 degrees of freedom

The model is able to explain 15.34% variation in Average_monthly_hours.

Therefore,

Average_monthly_hours = 110.873 + 0.1837* (Last_Evaluation2) + 0.2833 * (number_project2) + 0.0403 * (left2) + 0.03319 * (time_spend_company2)