EDA REPORT
CLEAN_DATA

Report Overview

This report was created for the EDA of ***clean_data*** data. It helps explore data to **understand the data and find scenarios for performing the analysis.**

Data by Nick Youngson CC BY-SA 3.0 Alpha Stock

Contents

Overview 2
- Data Structures 2
- Job Informations 2

Univariate Analysis 3
- Descriptive Statistics 3
  - Numerical Variables 3
  - Categorical Variables 6
- Normality Test 9

Bivariate Analysis 15
- Compare Numerical Variables 15
- Compare Categorical Variables 26

Multivariate Analysis 27
- Correlation Analysis 27
  - Correlation Coefficient Matrix 27
  - Correlation Plot 28

Overview

Data Structures

division	metrics	value
size	observations	1,000
size	variables	14
size	values	14,000
size	memory size (KB)	0
duplicated	duplicate observation	0
missing	complete observation	1,000
missing	missing observation	0
missing	missing variables	0
missing	missing values	0

division	metrics	value
data type	numerics	3
data type	integers	2
data type	factors/ordered	0
data type	characters	9
data type	Dates	0
data type	POSIXcts	0
data type	others	0

Table 1: Data structures and types

Job Informations

division	metrics	value
dataset	dataset	clean_data
dataset	dataset type	data.frame
dataset	target	not defied
job	samples	1,000 / 1,000 (100%)
job	created	2024-04-04 11:51:42.414756
job	created by	dlookr

Table 2: Job informations

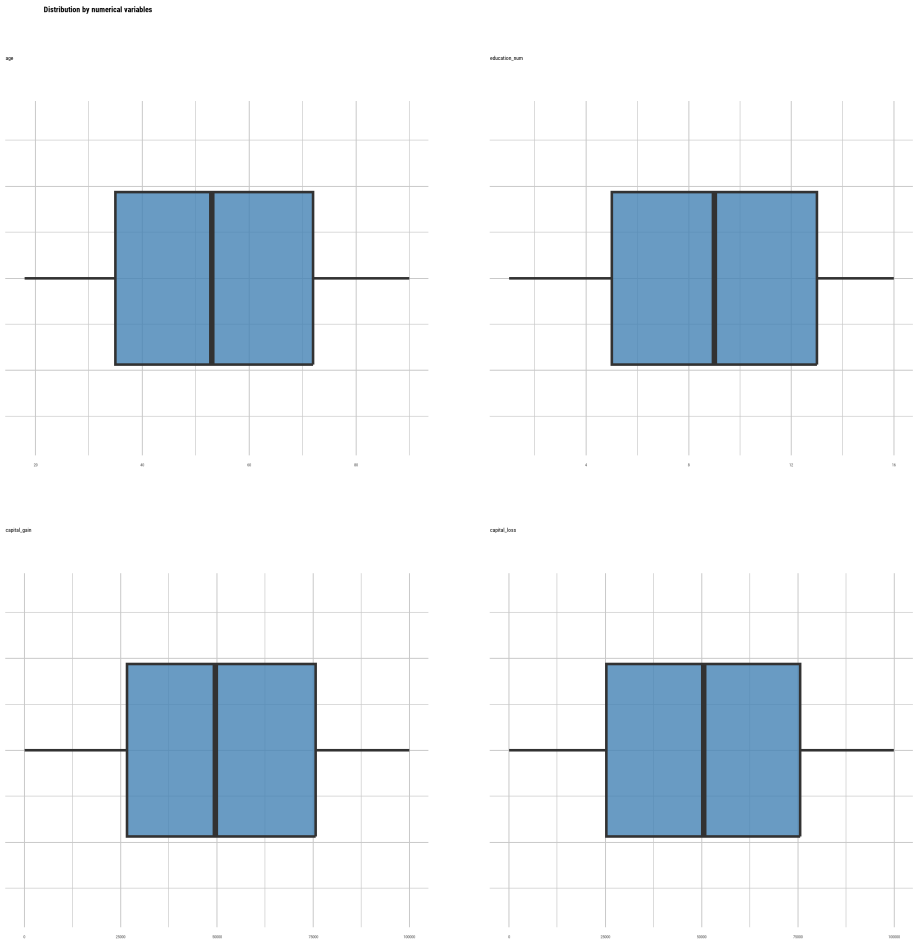
Univariate Analysis

Descriptive Statistics

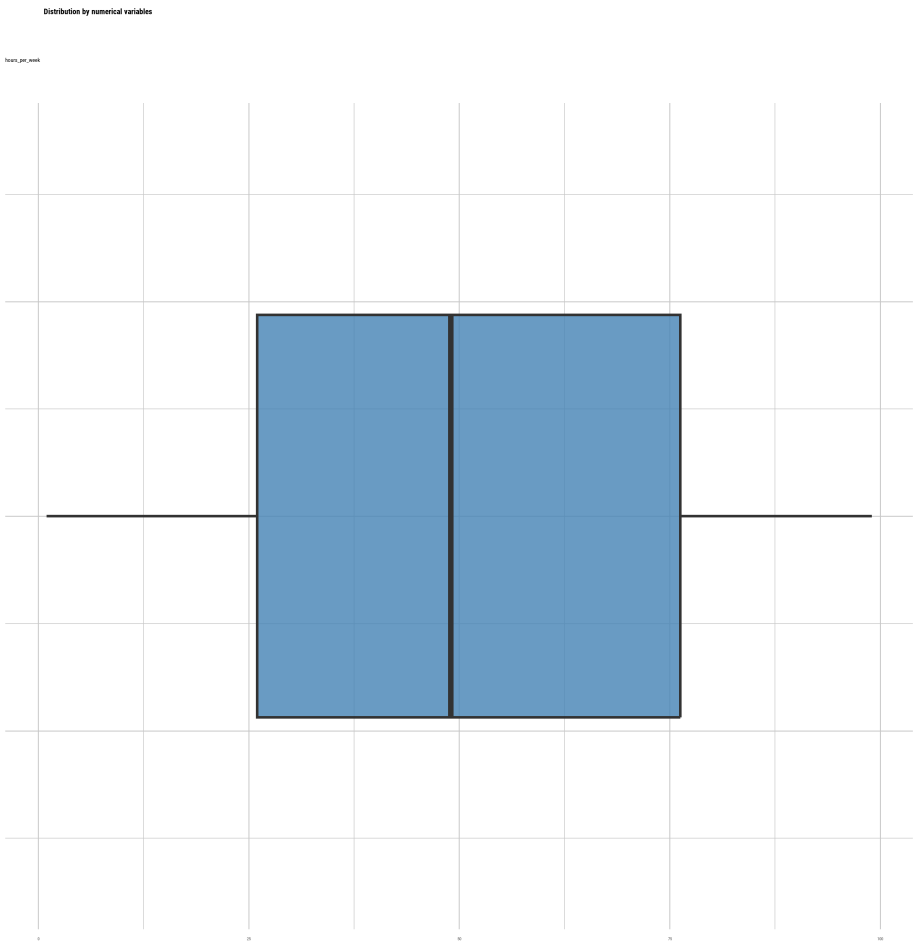
Numerical Variables

variables	missing	mean	sd	min	Q1	median	Q3	max
age	0	53.47	21.07	18	35.00	53	72.00	90
education_num	0	8.59	4.65	1	5.00	9	13.00	16
capital_gain	0	50,748.19	28,311.50	85	26,662.50	49,541	75,636.25	99,980
capital_loss	0	50,140.91	29,148.23	55	25,252.25	50,544	75,620.75	99,897
hours_per_week	0	50.19	28.84	1	26.00	49	76.25	99

Table 3: Descriptive statistics of numerical variables



variables	data types	distinct	skewness	kurtosis	zero	negative	outlier
age	numeric	73	0.02	-1.21	0	0	0
education_num	integer	16	-0.02	-1.22	0	0	0
capital_gain	numeric	944	0.00	-1.19	0	0	0
capital_loss	numeric	943	-0.03	-1.22	0	0	0



variables	data types	distinct	skewness	kurtosis	zero	negative	outlier
hours_per_week	integer	99	0.01	-1.2	0	0	0

Categorical Variables

variables	levels	observations	frequency	frequency(%)	rank
workclass	Local-gov	1,000	143	14.3	1
workclass	Federal-gov	1,000	141	14.1	2
workclass	Self-emp-not-inc	1,000	130	13.0	3
workclass	Private	1,000	127	12.7	4
workclass	Without-pay	1,000	126	12.6	5
workclass	State-gov	1,000	115	11.5	6
workclass	Never-worked	1,000	111	11.1	7
workclass	Self-emp-inc	1,000	107	10.7	8
education	Masters	1,000	79	7.9	1
education	1st-4th	1,000	74	7.4	2
education	9th	1,000	73	7.3	3
education	Assoc-voc	1,000	70	7.0	4
education	11th	1,000	68	6.8	5
education	5th-6th	1,000	68	6.8	5
education	Doctorate	1,000	66	6.6	7
education	Some-college	1,000	64	6.4	8
education	Bachelors	1,000	61	6.1	9
education	HS-grad	1,000	61	6.1	9
marital_status	Never-married	1,000	155	15.5	1
marital_status	Widowed	1,000	152	15.2	2
marital_status	Married-civ-spouse	1,000	149	14.9	3
marital_status	Married-spouse-absent	1,000	149	14.9	3
marital_status	Divorced	1,000	140	14.0	5
marital_status	Separated	1,000	139	13.9	6
marital_status	Married-AF-spouse	1,000	116	11.6	7

Table 4: Top rank levels of categorical variables

	variables	levels	observations	frequency	frequency(%)	rank
	variables	levels	observations	frequency	frequency(%)	rank
26	occupation	Handlers-cleaners	1,000	90	9.0	1
27	occupation	Machine-op-inspct	1,000	82	8.2	2
28	occupation	Armed-Forces	1,000	80	8.0	3
29	occupation	Sales	1,000	79	7.9	4
30	occupation	Prof-specialty	1,000	76	7.6	5
31	occupation	Exec-managerial	1,000	72	7.2	6
32	occupation	Protective-serv	1,000	72	7.2	6
33	occupation	Transport-moving	1,000	72	7.2	6
34	occupation	Craft-repair	1,000	66	6.6	9
35	occupation	Farming-fishing	1,000	65	6.5	10
36	relationship	Husband	1,000	181	18.1	1
37	relationship	Wife	1,000	173	17.3	2
38	relationship	Unmarried	1,000	166	16.6	3
39	relationship	Not-in-family	1,000	163	16.3	4
40	relationship	Own-child	1,000	160	16.0	5
41	relationship	Other-relative	1,000	157	15.7	6
42	race	White	1,000	213	21.3	1
43	race	Amer-Indian-Eskimo	1,000	210	21.0	2
44	race	Other	1,000	207	20.7	3
45	race	Black	1,000	191	19.1	4
46	race	Asian-Pac-Islander	1,000	179	17.9	5
47	sex	Female	1,000	507	50.7	1
48	sex	Male	1,000	493	49.3	2
49	native_country	Mexico	1,000	39	3.9	1
50	native_country	Yugoslavia	1,000	36	3.6	2
51	native_country	Greece	1,000	33	3.3	3
52	native_country	Outlying-US(Guam-USVI-etc)	1,000	32	3.2	4
53	native_country	Puerto-Rico	1,000	32	3.2	4

Table 4: Top rank levels of categorical variables (continued)

	variables	levels	observations	frequency	frequency(%)	rank
54	native_country	Trinadad&Tobago	1,000	32	3.2	4
55	native_country	India	1,000	31	3.1	7
56	native_country	Cuba	1,000	30	3.0	8
57	native_country	England	1,000	30	3.0	8
58	native_country	Canada	1,000	29	2.9	10
59	income	<=50K	1,000	737	73.7	1
60	income	>50K	1,000	263	26.3	2

Table 4: Top rank levels of categorical variables (continued)

The number of categorical(factor/ordered) variables is 0.

Normality Test

described_variables	min	Q1	median	Q3	max	skewness	kurtosis	balance
age	18	35.0	53	72.0	90	0	-1.2	Balanced
education_num	1	5.0	9	13.0	16	0	-1.2	Balanced
capital_gain	85	26662.5	49541	75636.2	99980	0	-1.2	Balanced
capital_loss	55	25252.2	50544	75620.8	99897	0	-1.2	Balanced
hours_per_week	1	26.0	49	76.2	99	0	-1.2	Balanced

Table 5: Descriptive statistics of numerical variables

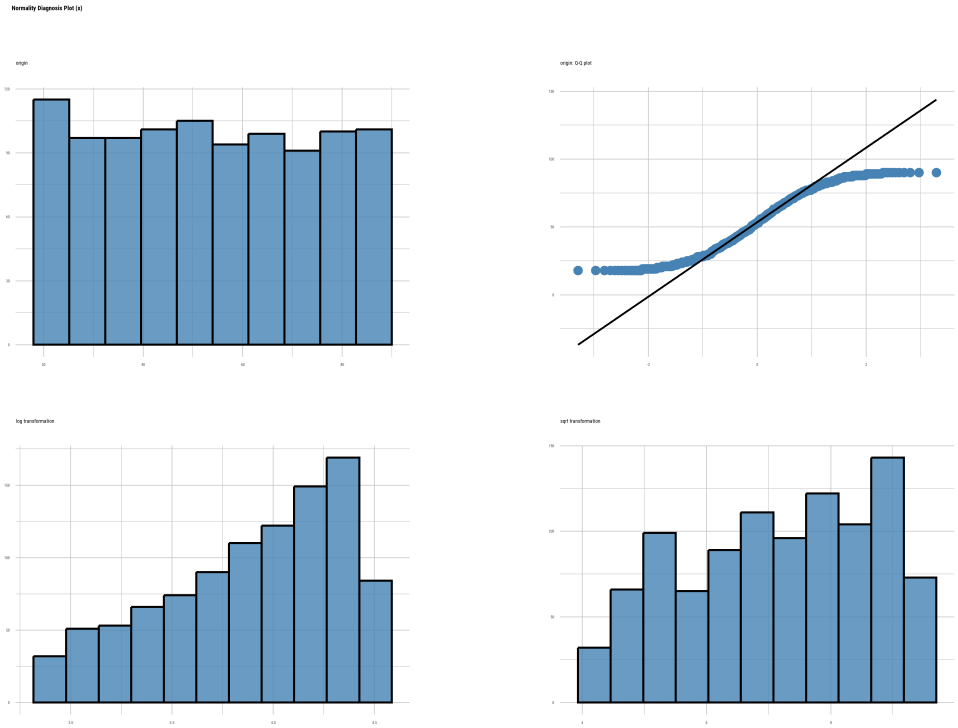
age

statistic	p_value	remark
0.95365	3.1749e-17	No sample

Table 6: Shapiro-Wilk normality test

type	skewness	kurtosis
original	0.0209	1.7925
log transformation	-0.5370	2.2005
sqrt transformation	-0.2442	1.8981

Table 6: skewness and kurtosis



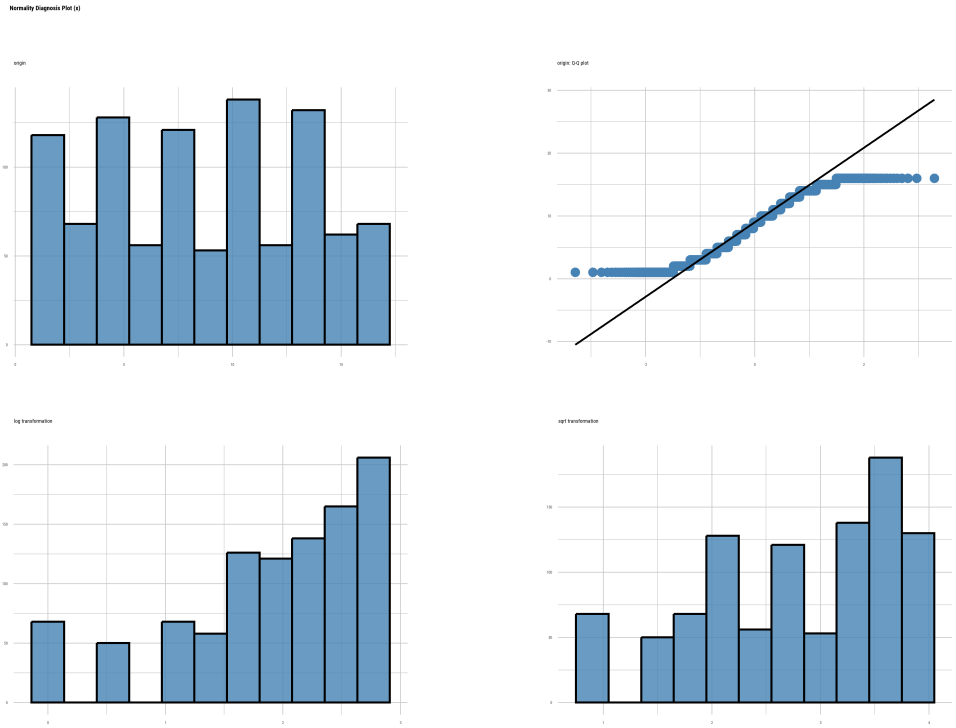
education_num

statistic	p_value	remark
0.94444	6.3912e-19	No sample

Table 6: Shapiro-Wilk normality test

type	skewness	kurtosis
original	-0.0176	1.7799
log transformation	-1.0767	3.3499
sqrt transformation	-0.4665	2.1511

Table 6: skewness and kurtosis



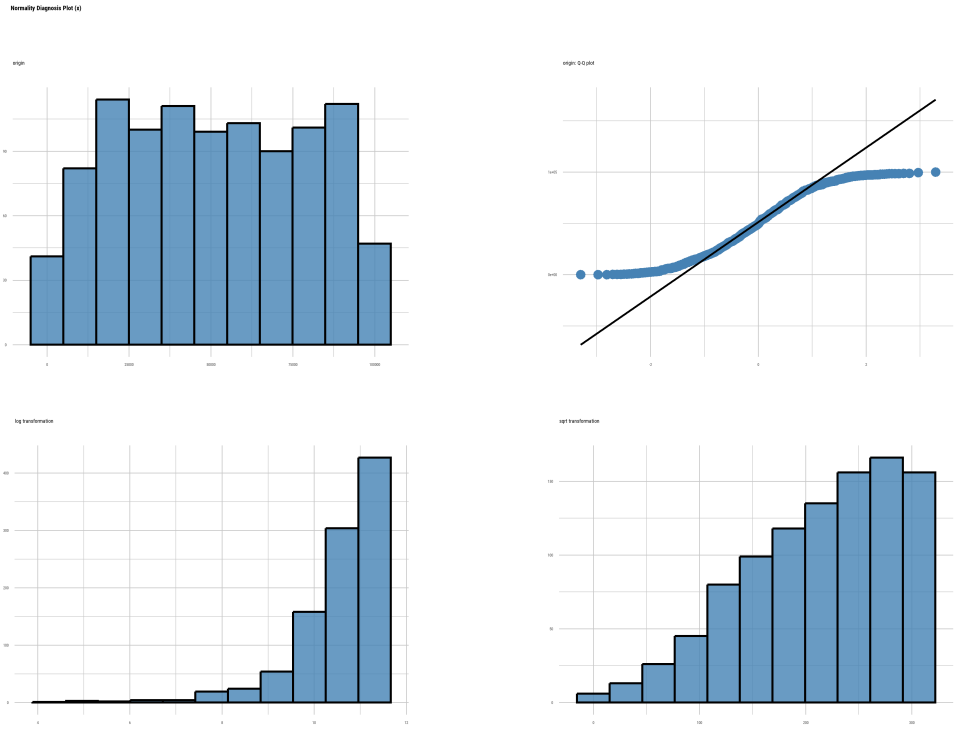
capital_gain

statistic	p_value	remark
0.95645	1.1591e-16	No sample

Table 6: Shapiro-Wilk normality test

type	skewness	kurtosis
original	-0.0033	1.8132
log transformation	-2.2860	10.8044
sqrt transformation	-0.5926	2.5577

Table 6: skewness and kurtosis



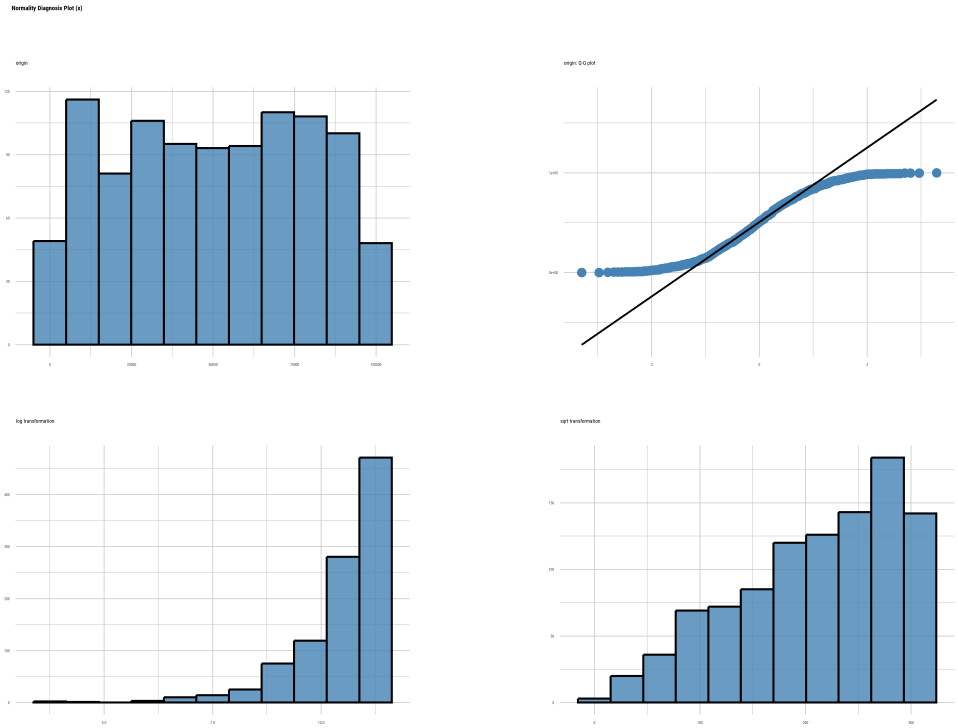
capital_loss

statistic	p_value	remark
0.95199	1.5074e-17	No sample

Table 6: Shapiro-Wilk normality test

type	skewness	kurtosis
original	-0.0301	1.7758
log transformation	-2.0619	9.1609
sqrt transformation	-0.5861	2.3824

Table 6: skewness and kurtosis



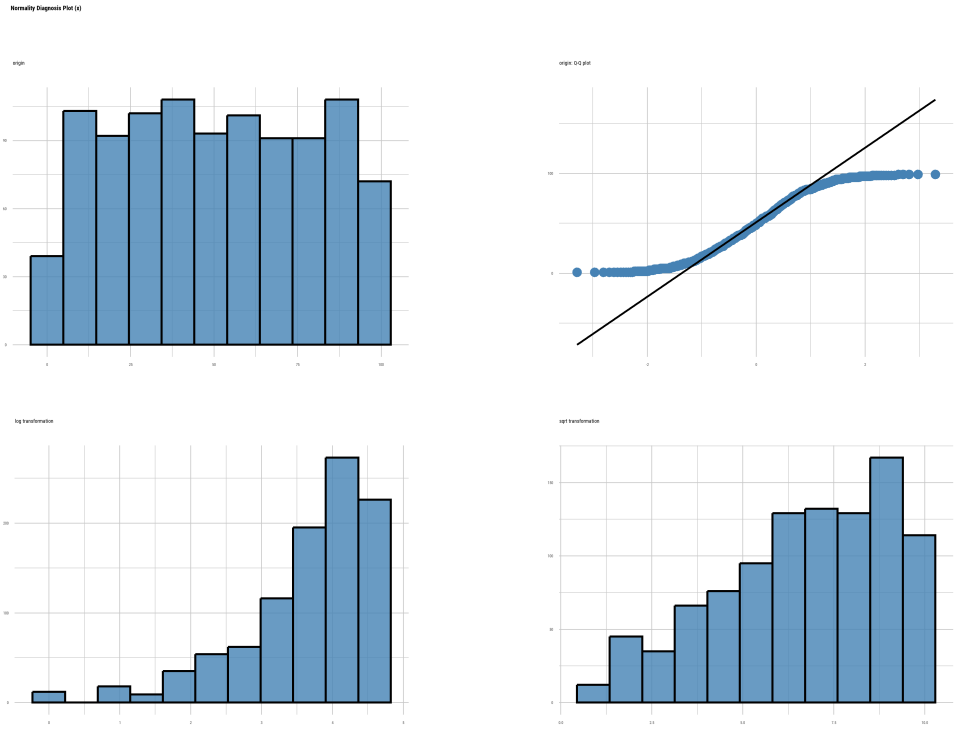
hours_per_week

statistic	p_value	remark
0.95251	1.9032e-17	No sample

Table 6: Shapiro-Wilk normality test

type	skewness	kurtosis
original	0.0087	1.7969
log transformation	-1.5655	5.4855
sqrt transformation	-0.5434	2.3569

Table 6: skewness and kurtosis



Bivariate Analysis

Compare Numerical Variables

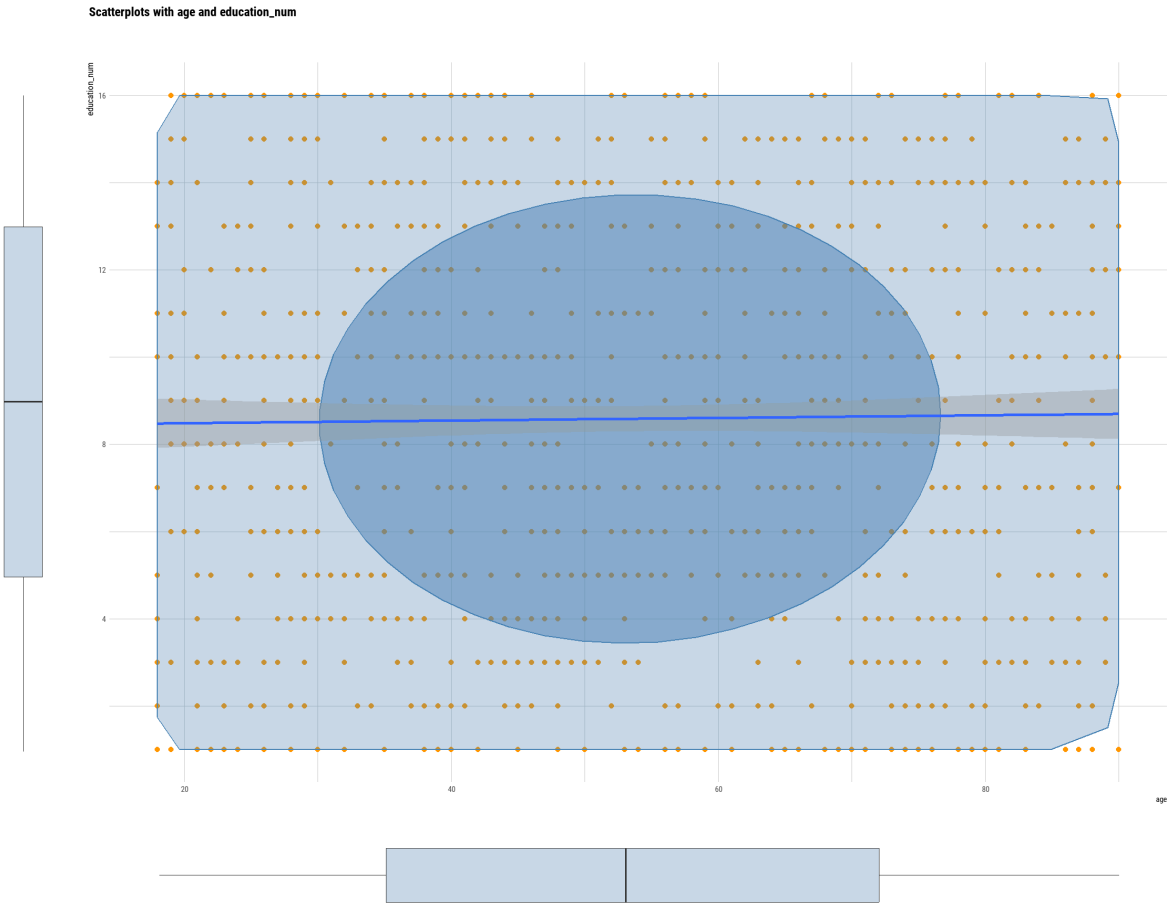
first variable	second variable	correlation coefficient
age	education_num	0.01369
age	capital_gain	-0.03187
age	capital_loss	0.01652
age	hours_per_week	0.00489
education_num	capital_gain	-0.00146
education_num	capital_loss	0.05191
education_num	hours_per_week	-0.03335
capital_gain	capital_loss	-0.08995
capital_gain	hours_per_week	-0.01275
capital_loss	hours_per_week	0.00618

Table 7: Correlation coefficient

‘age’ vs ‘education_num’

first variable	second variable	r.squared	adj.r.squared	sigma	statistic	p.value	df
age	education_num	0.0001875	-0.0008144	21.0831	0.1871179	0.6654197	1

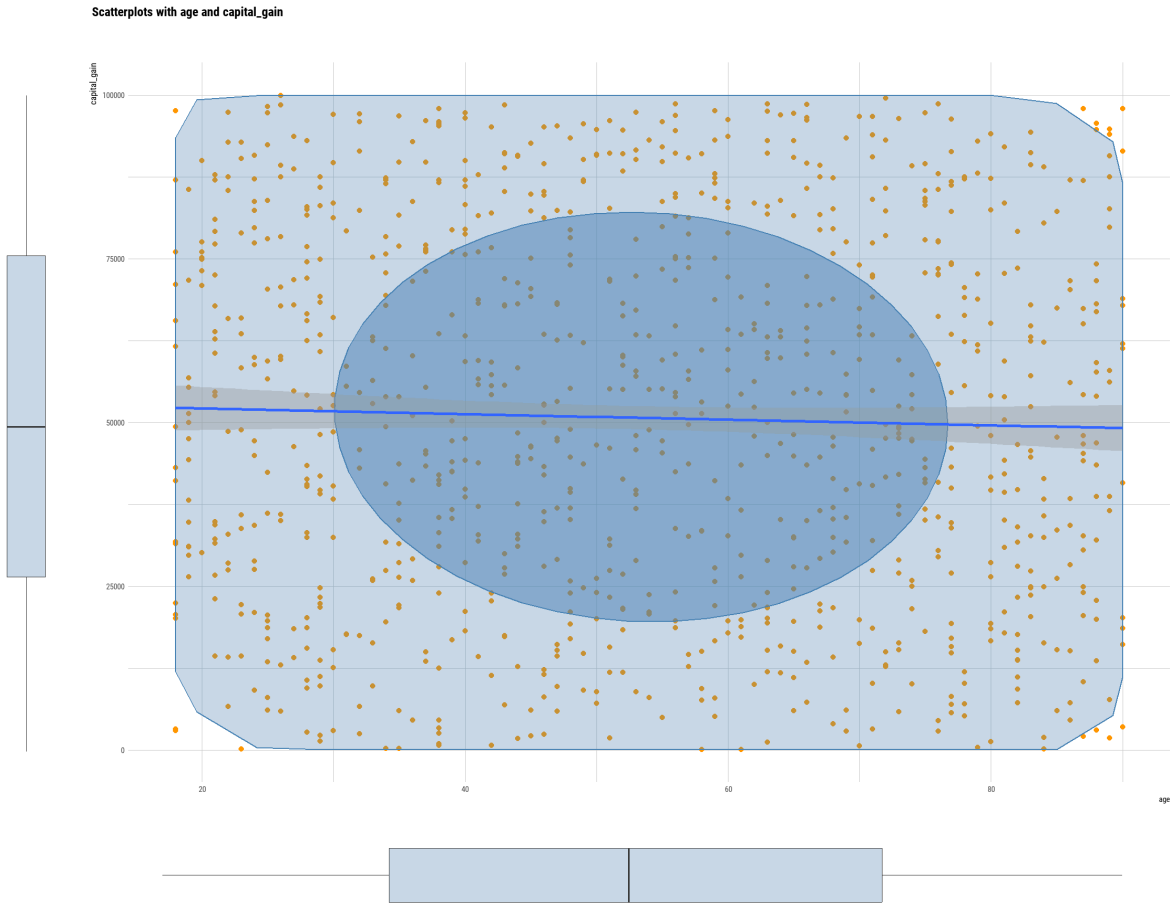
Table 7: Summary of linear model



‘age’ vs ‘capital_gain’

first variable	second variable	r.squared	adj.r.squared	sigma	statistic	p.value	df
age	capital_gain	0.0010154	1.44e-05	21.07436	1.014388	0.314098	1

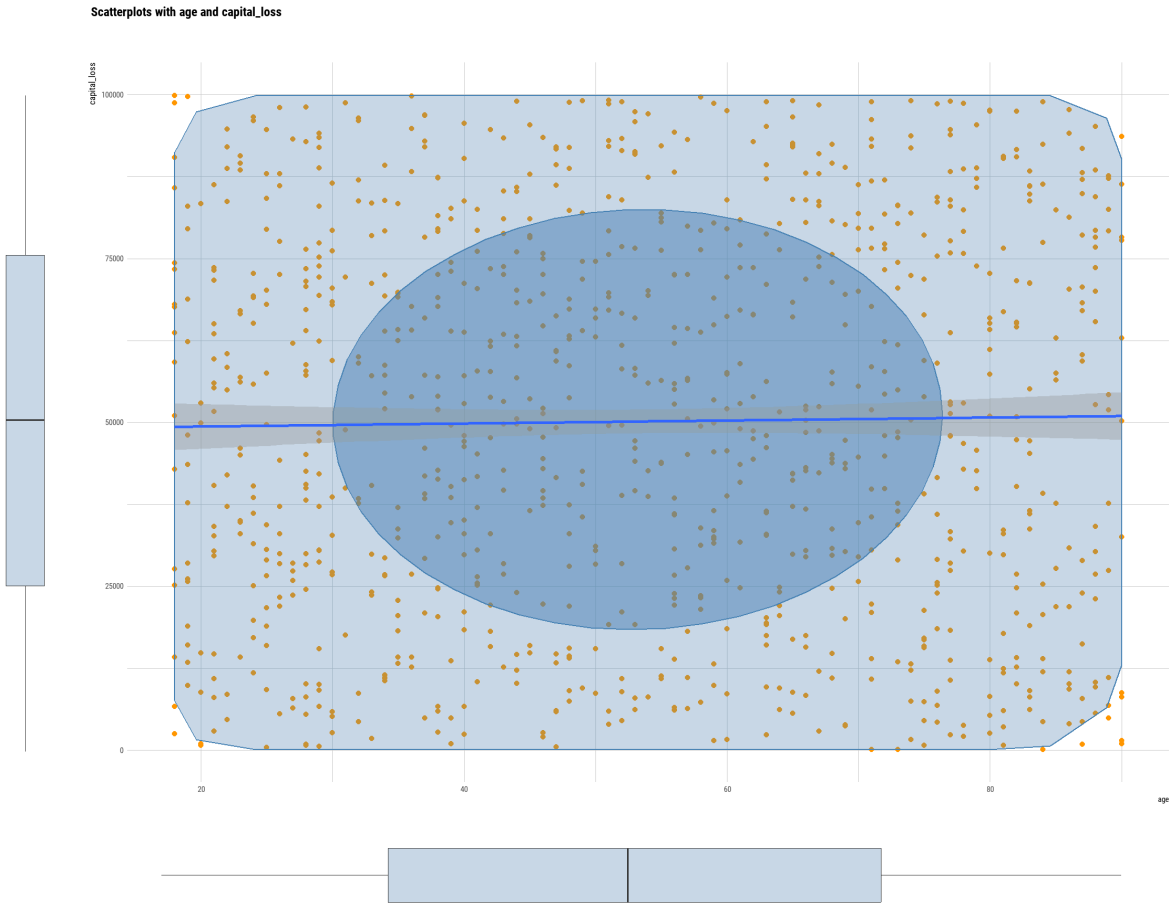
Table 7: Summary of linear model



‘age’ vs ‘capital_loss’

first variable	second variable	r.squared	adj.r.squared	sigma	statistic	p.value	df
age	capital_loss	0.000273	-0.0007287	21.08219	0.2725492	0.6017427	1

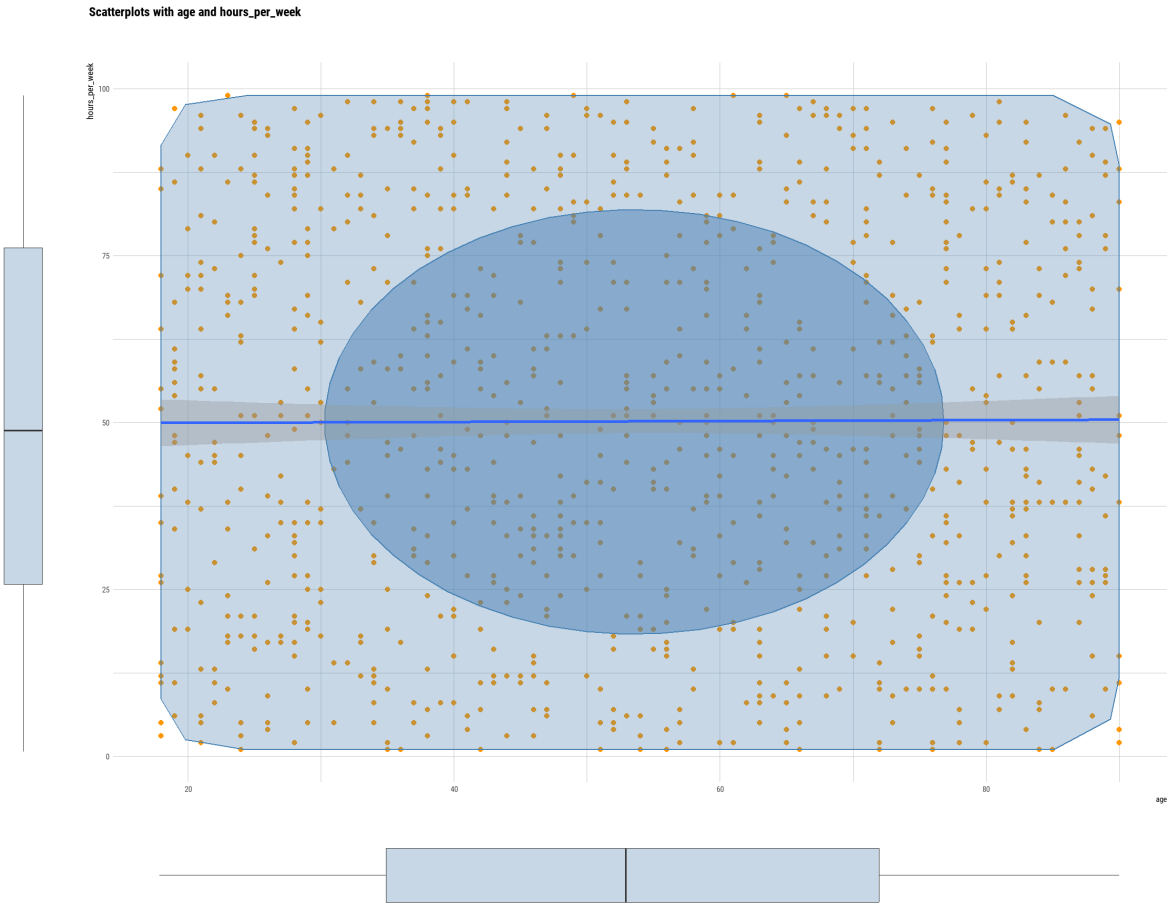
Table 7: Summary of linear model



‘age’ vs ‘hours_per_week’

first variable	second variable	r.squared	adj.r.squared	sigma	statistic	p.value	df
age	hours_per_week	2.39e-05	-0.0009781	21.08482	0.0238452	0.8773106	1

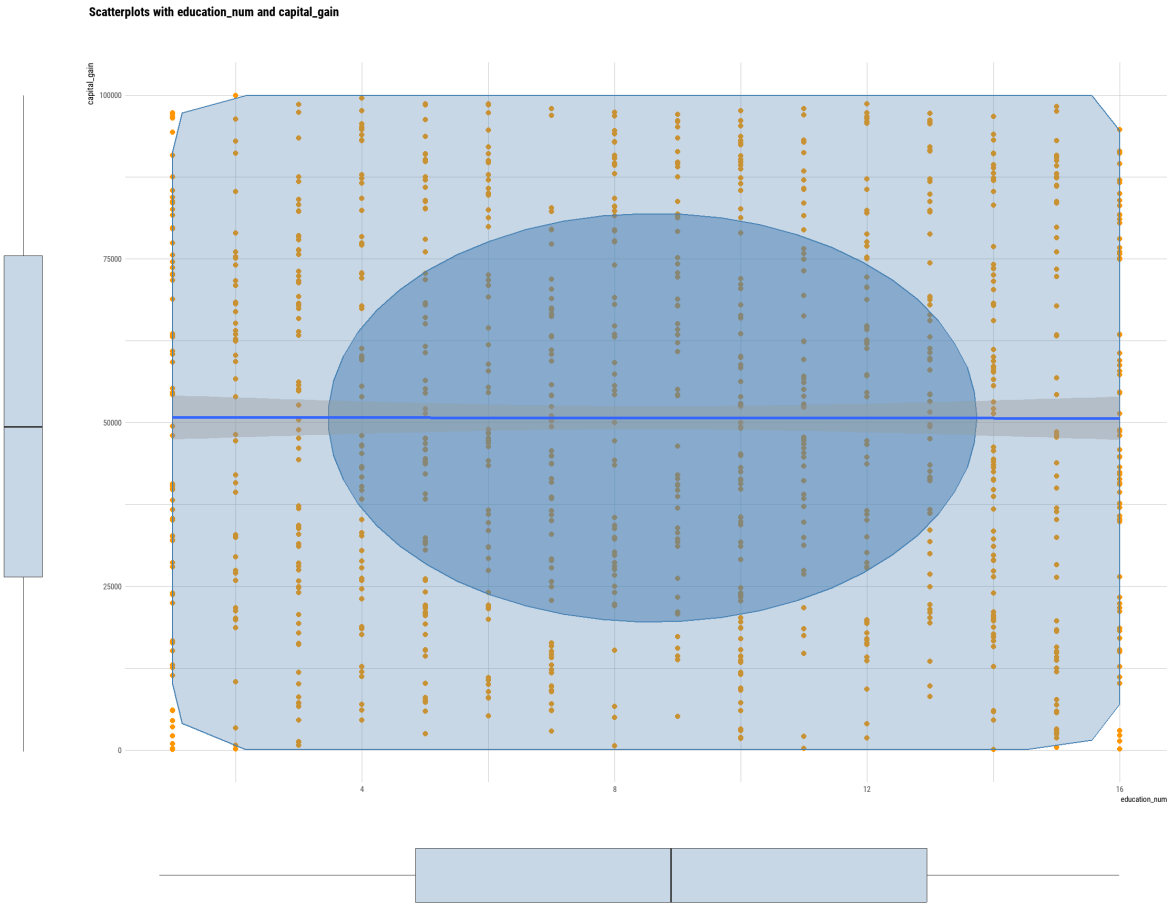
Table 7: Summary of linear model



‘education_num’ vs ‘capital_gain’

first variable	second variable	r.squared	adj.r.squared	sigma	statistic	p.value	df
education_num	capital_gain	2.1e-06	-0.0009999	4.65486	0.0021145	0.9633323	1

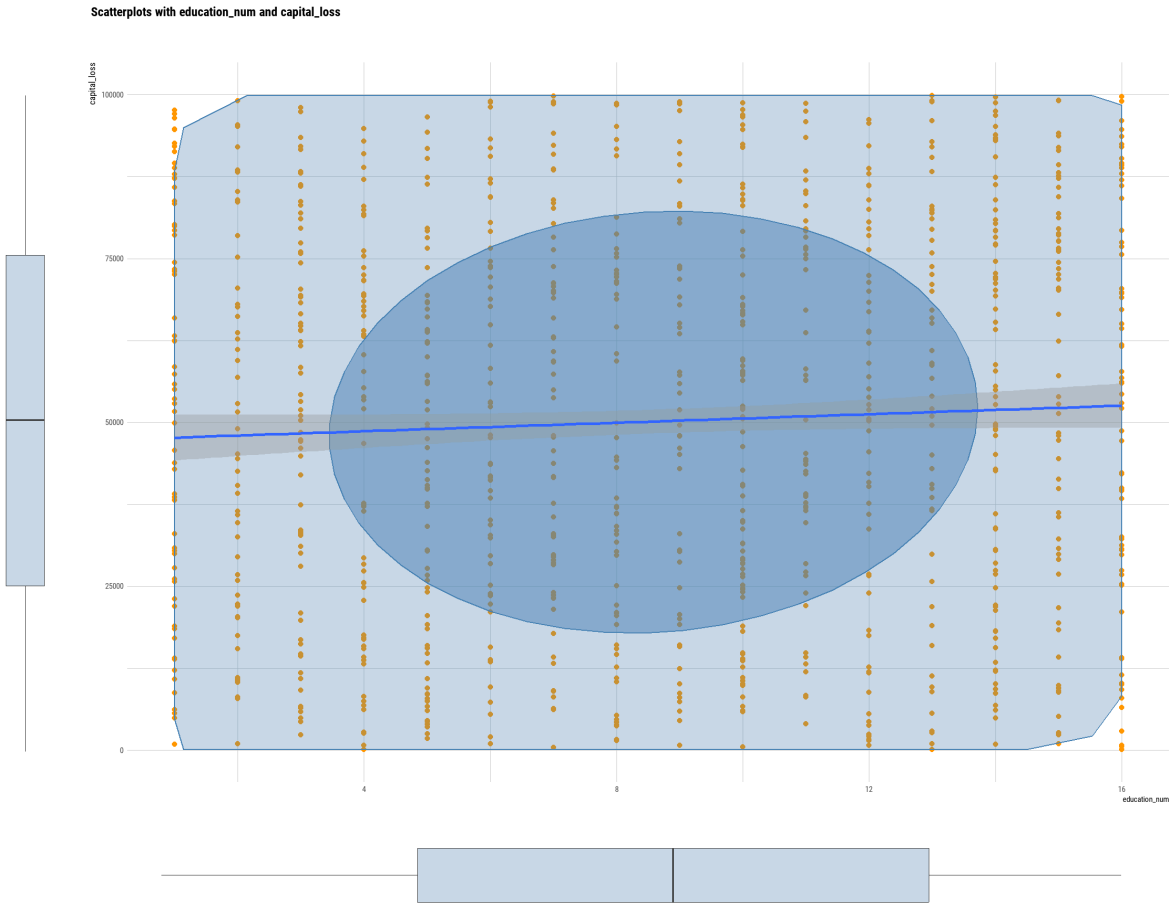
Table 7: Summary of linear model



‘education_num’ vs ‘capital_loss’

first variable	second variable	r.squared	adj.r.squared	sigma	statistic	p.value	df
education_num	capital_loss	0.0026948	0.0016955	4.648589	2.69664	0.100875	1

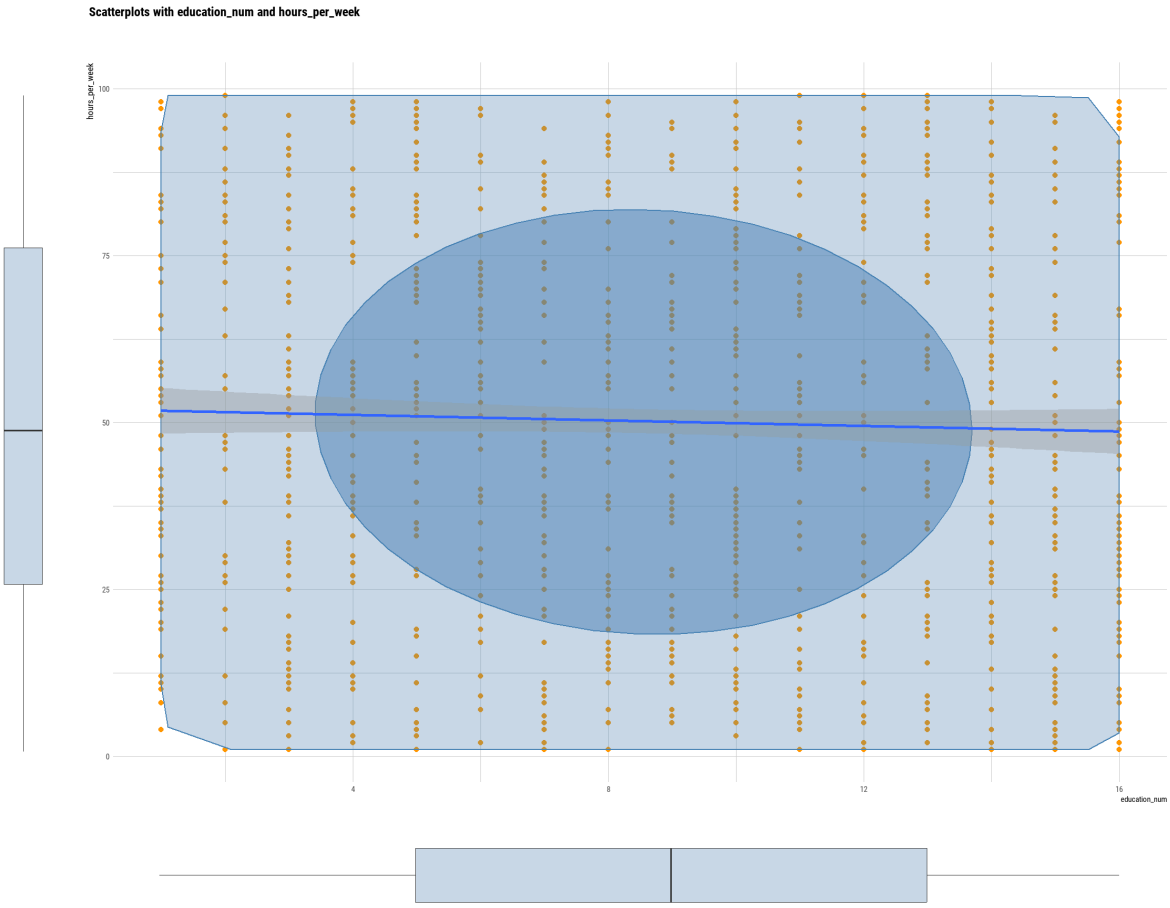
Table 7: Summary of linear model



‘education_num’ vs ‘hours_per_week’

first variable	second variable	r.squared	adj.r.squared	sigma	statistic	p.value	df
education_num	hours_per_week	0.0011124	0.0001115	4.652275	1.111405	0.2920319	1

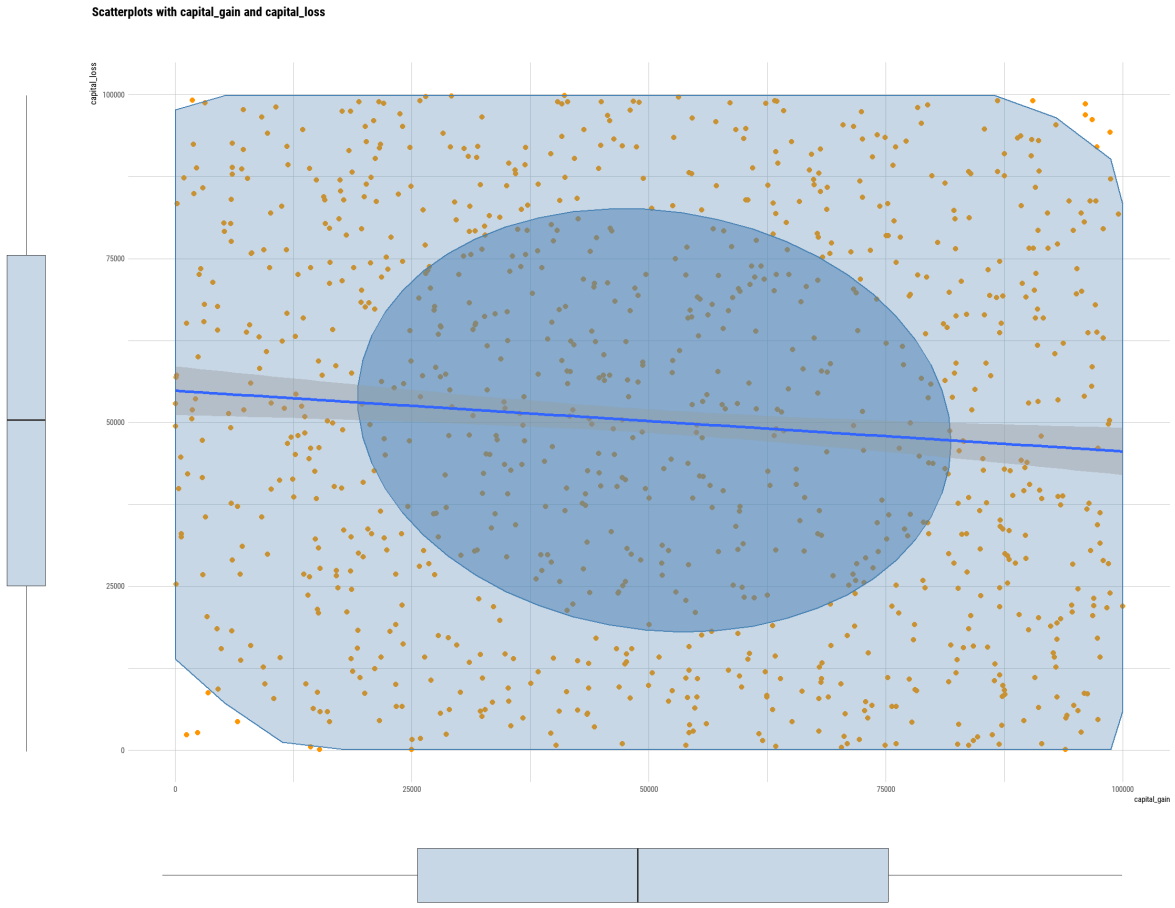
Table 7: Summary of linear model



‘capital_gain’ vs ‘capital_loss’

first variable	second variable	r.squared	adj.r.squared	sigma	statistic	p.value	df
capital_gain	capital_loss	0.0080918	0.0070979	28210.85	8.141512	0.0044157	1

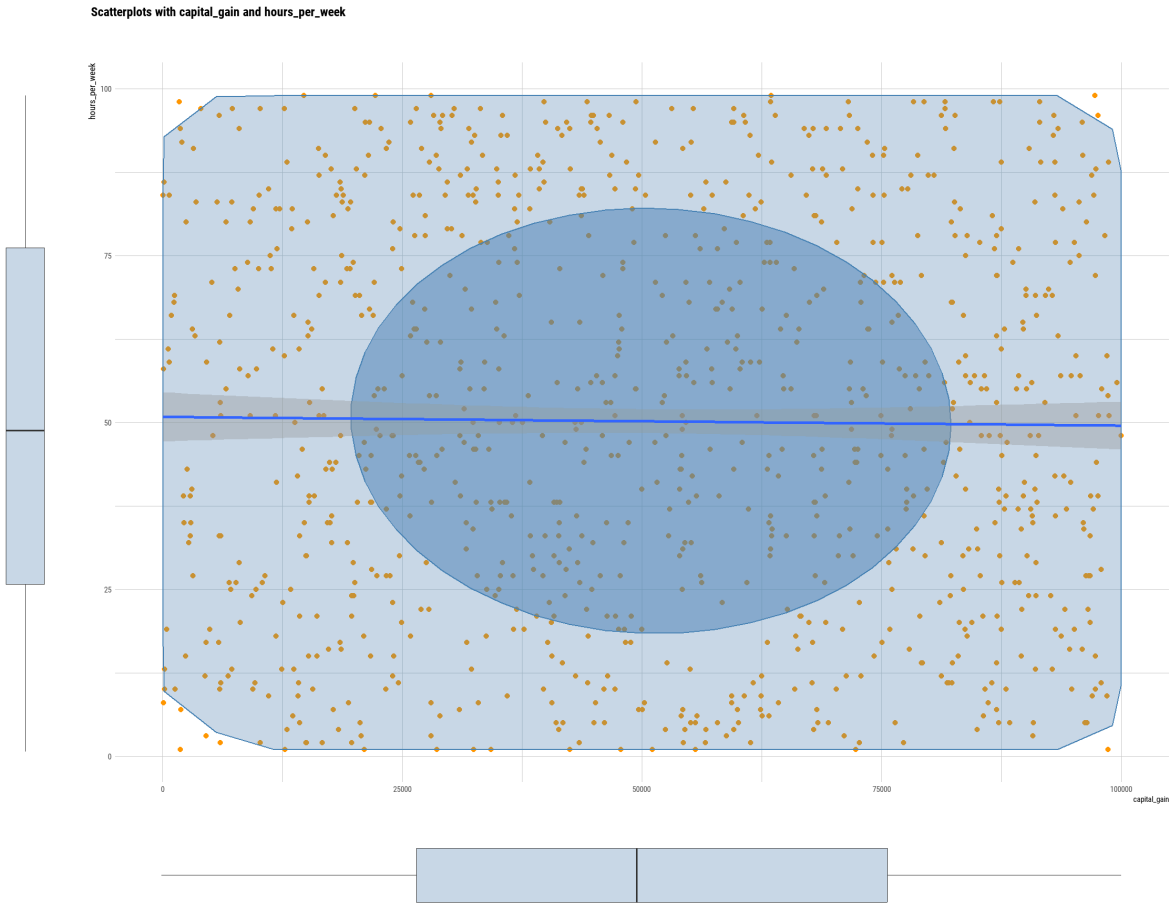
Table 7: Summary of linear model



‘capital_gain’ vs ‘hours_per_week’

first variable	second variable	r.squared	adj.r.squared	sigma	statistic	p.value	df
capital_gain	hours_per_week	0.0001626	-0.0008392	28323.38	0.1622986	0.6871353	1

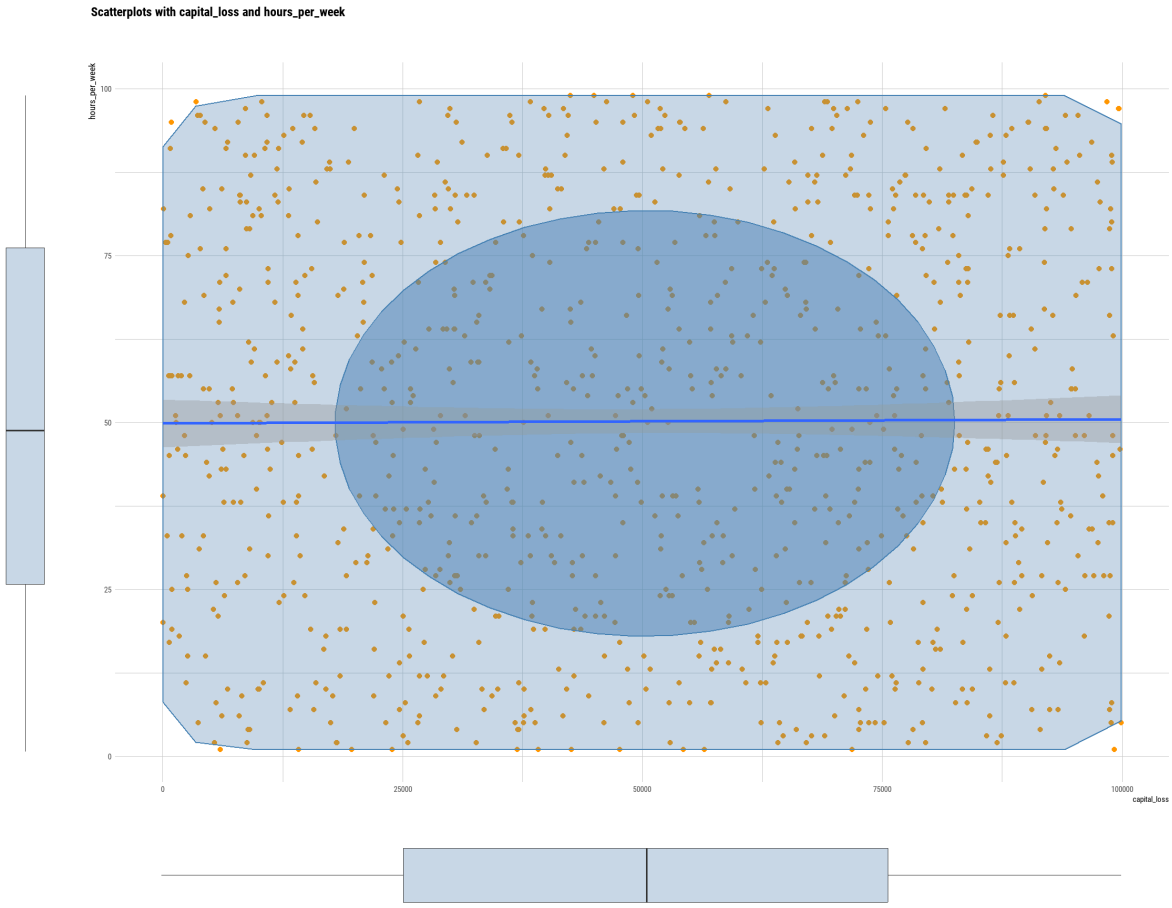
Table 7: Summary of linear model



‘capital_loss’ vs ‘hours_per_week’

first variable	second variable	r.squared	adj.r.squared	sigma	statistic	p.value	df
capital_loss	hours_per_week	3.82e-05	-0.0009637	29162.27	0.0381702	0.8451416	1

Table 7: Summary of linear model



Compare Categorical Variables

The number of categorical variables is less than 2.

Multivariate Analysis

Correlation Analysis

Correlation Coefficient Matrix

	second variable				
first variable	**age**	**education_num**	**capital_gain**	**capital_loss**	**hours_per_week**
age	NA	0.014	-0.032	0.017	0.005
education_num	0.014	NA	-0.001	0.052	-0.033
capital_gain	-0.032	-0.001	NA	-0.090	-0.013
capital_loss	0.017	0.052	-0.090	NA	0.006
hours_per_week	0.005	-0.033	-0.013	0.006	NA

Table 8: Matrix table of correlation coefficient

Correlation Plot