EDA REPORT
SAMPLE_DATA_WITH_NA

Report Overview

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

Contents

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	525
missing	missing observation	475
missing	missing variables	12
missing	missing values	600

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	sample_data_with_NA
dataset	dataset type	data.frame
dataset	target	not defied
job	samples	1,000 / 1,000 (100%)
job	created	2024-04-04 11:49:48.098165
job	created by	dlookr

Table 2: Job informations

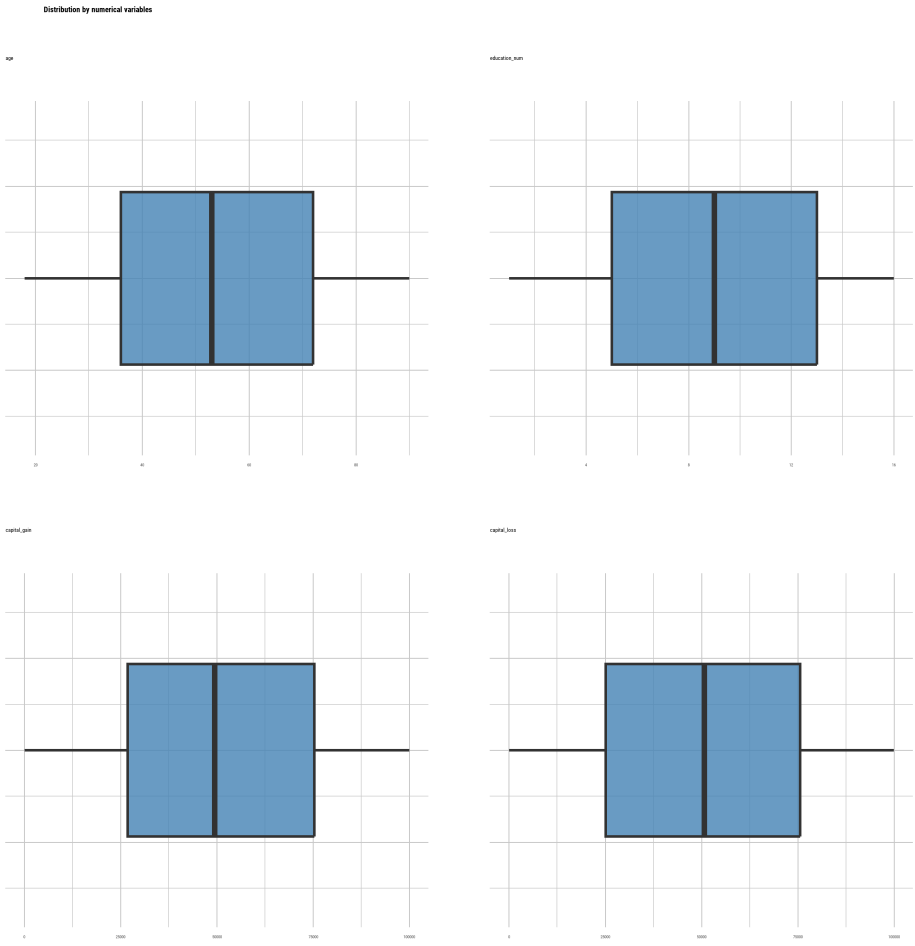
Univariate Analysis

Descriptive Statistics

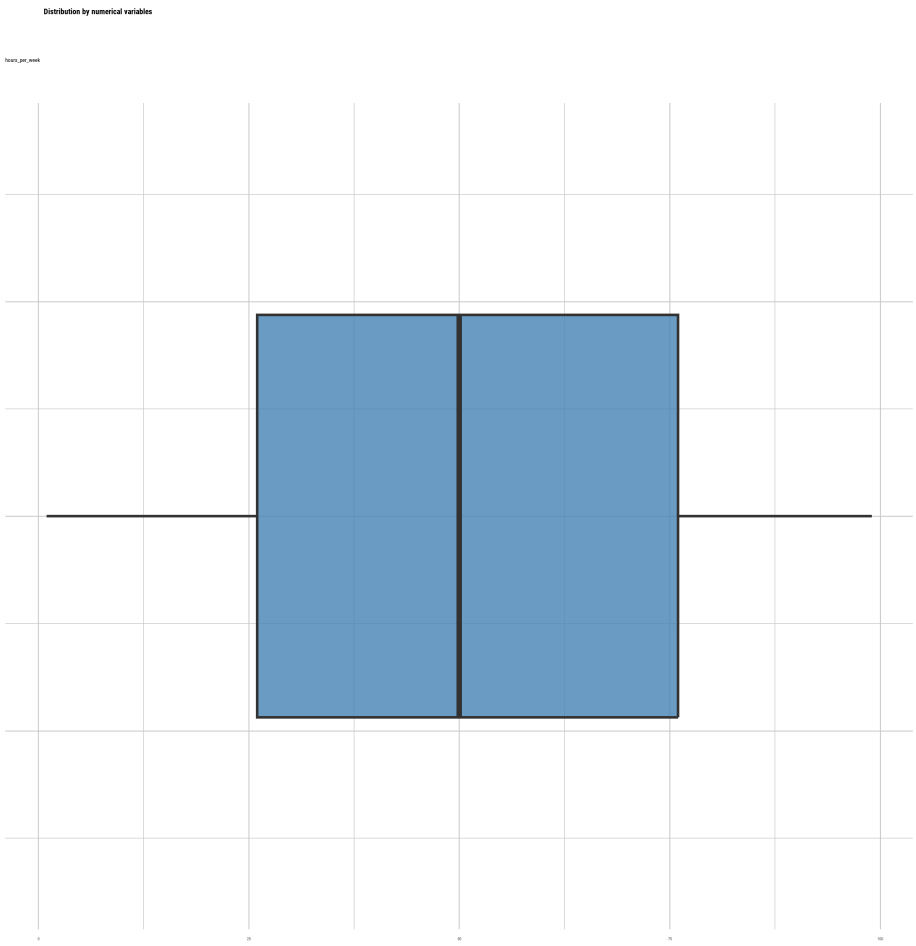
Numerical Variables

variables	missing	mean	sd	min	Q1	median	Q3	max
age	50	53.81	20.99	18	36.00	53.0	72.00	90
education_num	0	8.59	4.65	1	5.00	9.0	13.00	16
capital_gain	50	50,800.46	28,214.93	85	26,820.00	49,397.5	75,315.50	99,980
capital_loss	50	50,171.11	29,072.33	55	25,106.75	50,678.0	75,626.25	99,897
hours_per_week	50	50.25	28.86	1	26.00	50.0	76.00	99

Table 3: Descriptive statistics of numerical variables



variables	data types	distinct	skewness	kurtosis	zero	negative	outlier
age	numeric	74	0.00	-1.20	0	0	0
education_num	integer	16	-0.02	-1.22	0	0	0
capital_gain	numeric	945	0.01	-1.18	0	0	0
capital_loss	numeric	944	-0.03	-1.22	0	0	0



variables	data types	distinct	skewness	kurtosis	zero	negative	outlier
hours_per_week	integer	100	-0.01	-1.2	0	0	0

Categorical Variables

variables	levels	observations	frequency	frequency(%)	rank
workclass	Federal-gov	1,000	135	13.5	1
workclass	Local-gov	1,000	129	12.9	2
workclass	Without-pay	1,000	124	12.4	3
workclass	Self-emp-not-inc	1,000	123	12.3	4
workclass	Private	1,000	121	12.1	5
workclass	State-gov	1,000	109	10.9	6
workclass	Never-worked	1,000	107	10.7	7
workclass	Self-emp-inc	1,000	102	10.2	8
workclass	NA	1,000	50	5.0	9
education	Masters	1,000	74	7.4	1
education	1st-4th	1,000	69	6.9	2
education	Assoc-voc	1,000	69	6.9	2
education	9th	1,000	68	6.8	4
education	5th-6th	1,000	64	6.4	5
education	Doctorate	1,000	63	6.3	6
education	11th	1,000	61	6.1	7
education	Bachelors	1,000	60	6.0	8
education	Some-college	1,000	60	6.0	8
education	HS-grad	1,000	59	5.9	10
marital_status	Never-married	1,000	148	14.8	1
marital_status	Married-spouse-absent	1,000	141	14.1	2
marital_status	Widowed	1,000	141	14.1	2
marital_status	Married-civ-spouse	1,000	139	13.9	4
marital_status	Divorced	1,000	136	13.6	5
marital_status	Separated	1,000	132	13.2	6

Table 4: Top rank levels of categorical variables

	variables	levels	observations	frequency	frequency(%)	rank
	variables	levels	observations	frequency	frequency(%)	rank
26	marital_status	Married-AF-spouse	1,000	113	11.3	7
27	marital_status	NA	1,000	50	5.0	8
28	occupation	Handlers-cleaners	1,000	83	8.3	1
29	occupation	Machine-op-inspct	1,000	77	7.7	2
30	occupation	Armed-Forces	1,000	76	7.6	3
31	occupation	Sales	1,000	75	7.5	4
32	occupation	Prof-specialty	1,000	70	7.0	5
33	occupation	Protective-serv	1,000	69	6.9	6
34	occupation	Exec-managerial	1,000	68	6.8	7
35	occupation	Transport-moving	1,000	67	6.7	8
36	occupation	Craft-repair	1,000	63	6.3	9
37	occupation	Farming-fishing	1,000	63	6.3	9
38	relationship	Husband	1,000	175	17.5	1
39	relationship	Wife	1,000	160	16.0	2
40	relationship	Not-in-family	1,000	156	15.6	3
41	relationship	Unmarried	1,000	155	15.5	4
42	relationship	Own-child	1,000	153	15.3	5
43	relationship	Other-relative	1,000	151	15.1	6
44	relationship	NA	1,000	50	5.0	7
45	race	White	1,000	205	20.5	1
46	race	Amer-Indian-Eskimo	1,000	199	19.9	2
47	race	Other	1,000	199	19.9	2
48	race	Black	1,000	177	17.7	4
49	race	Asian-Pac-Islander	1,000	170	17.0	5
50	race	NA	1,000	50	5.0	6
51	sex	Female	1,000	481	48.1	1
52	sex	Male	1,000	469	46.9	2
53	sex	NA	1,000	50	5.0	3

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

	variables	levels	observations	frequency	frequency(%)	rank
54	native_country	NA	1,000	50	5.0	1
55	native_country	Mexico	1,000	37	3.7	2
56	native_country	Yugoslavia	1,000	33	3.3	3
57	native_country	Outlying-US(Guam-USVI-etc)	1,000	32	3.2	4
58	native_country	Greece	1,000	30	3.0	5
59	native_country	Puerto-Rico	1,000	30	3.0	5
60	native_country	Trinadad&Tobago	1,000	30	3.0	5
61	native_country	England	1,000	29	2.9	8
62	native_country	India	1,000	29	2.9	8
63	native_country	Canada	1,000	28	2.8	10
64	income	<=50K	1,000	737	73.7	1
65	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	36.0	53.0	72.0	90	0	-1.2	Balanced
education_num	1	5.0	9.0	13.0	16	0	-1.2	Balanced
capital_gain	85	26820.0	49397.5	75315.5	99980	0	-1.2	Balanced
capital_loss	55	25106.8	50678.0	75626.2	99897	0	-1.2	Balanced
hours_per_week	1	26.0	50.0	76.0	99	0	-1.2	Balanced

Table 5: Descriptive statistics of numerical variables

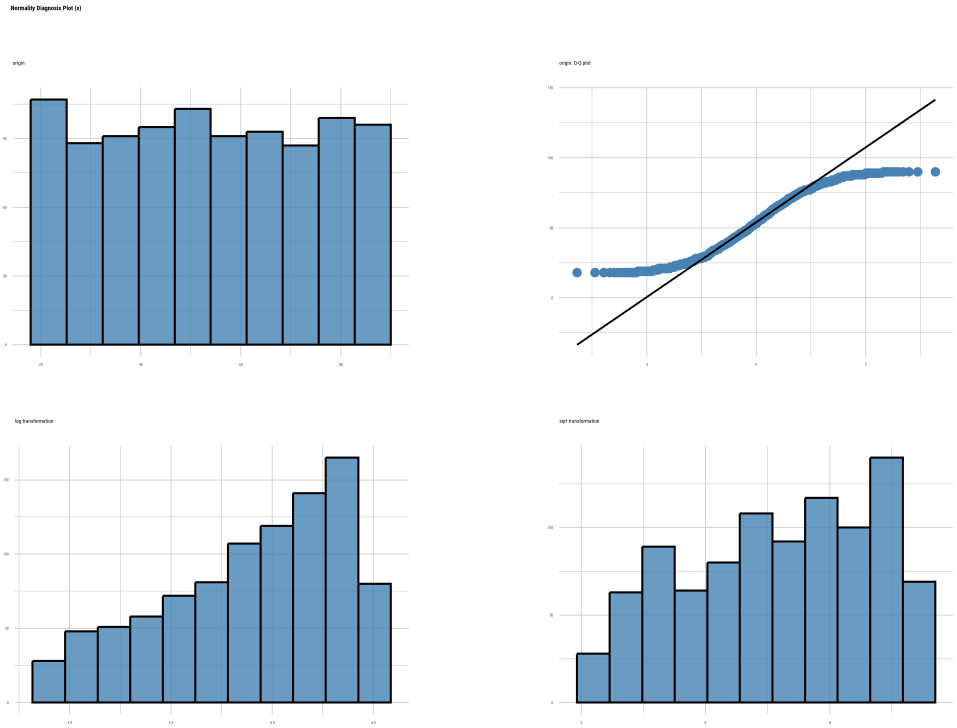
age

statistic	p_value	remark
0.95436	1.332e-16	No sample

Table 6: Shapiro-Wilk normality test

type	skewness	kurtosis
original	-0.0004	1.7982
log transformation	-0.5595	2.2335
sqrt transformation	-0.2661	1.9171

Table 6: skewness and kurtosis



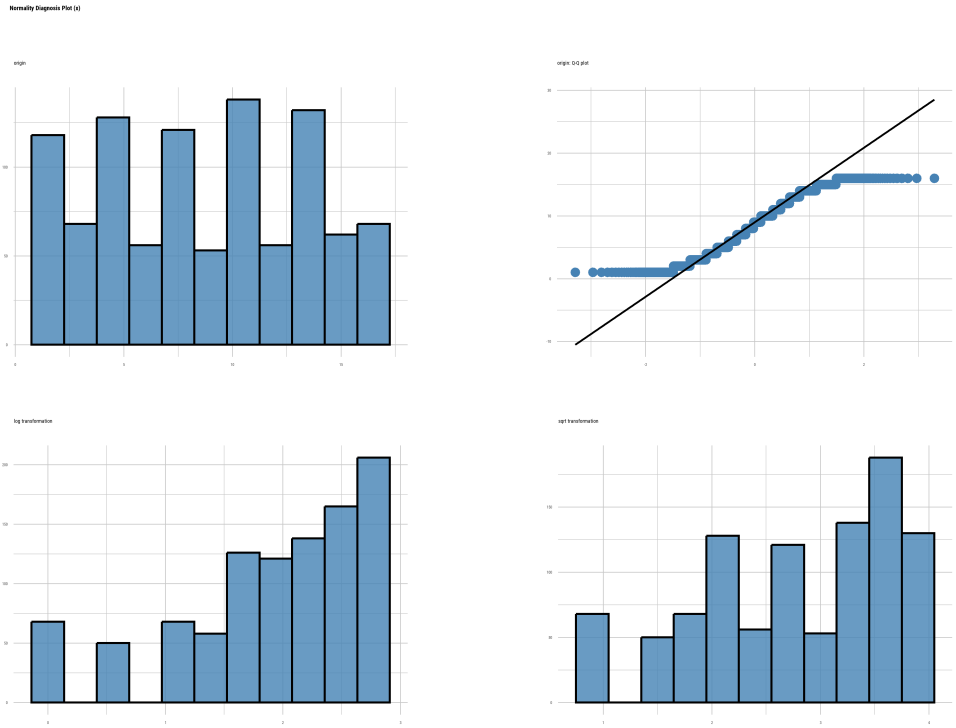
education_num

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



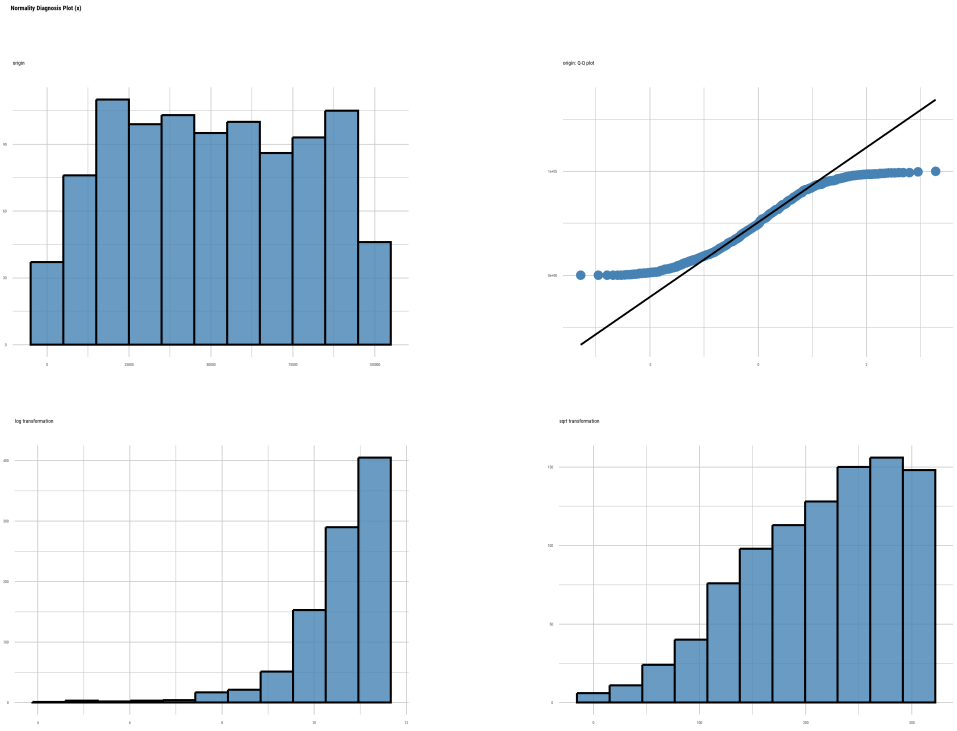
capital_gain

statistic	p_value	remark
0.95704	4.5716e-16	No sample

Table 6: Shapiro-Wilk normality test

type	skewness	kurtosis
original	0.0091	1.8190
log transformation	-2.3429	11.3534
sqrt transformation	-0.5831	2.5722

Table 6: skewness and kurtosis



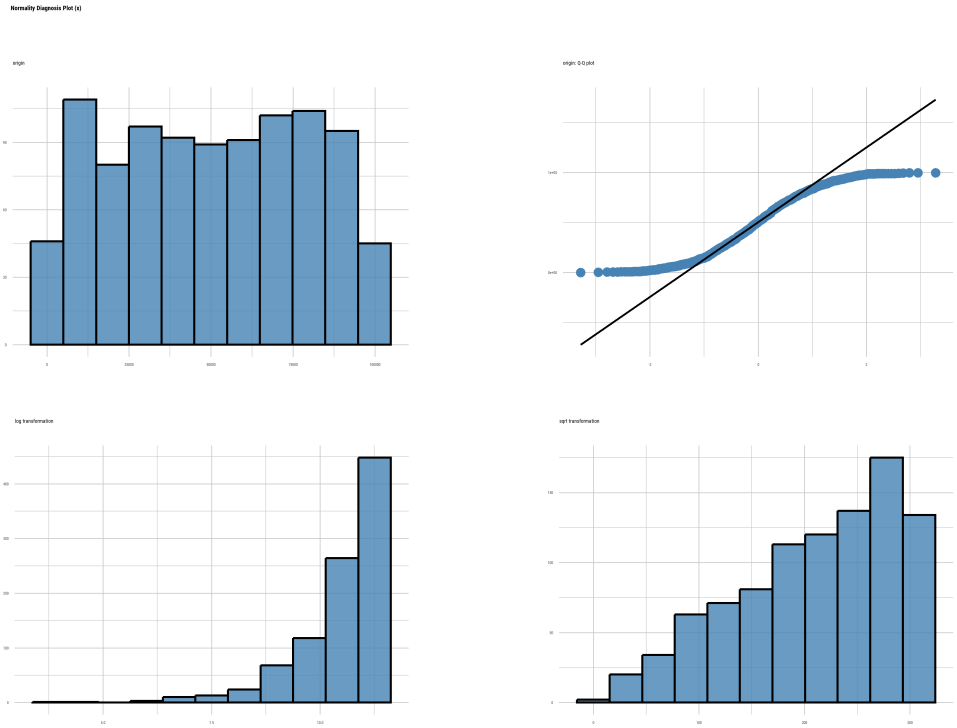
capital_loss

statistic	p_value	remark
0.95245	5.7317e-17	No sample

Table 6: Shapiro-Wilk normality test

type	skewness	kurtosis
original	-0.0330	1.7792
log transformation	-1.9734	8.4311
sqrt transformation	-0.5870	2.3829

Table 6: skewness and kurtosis



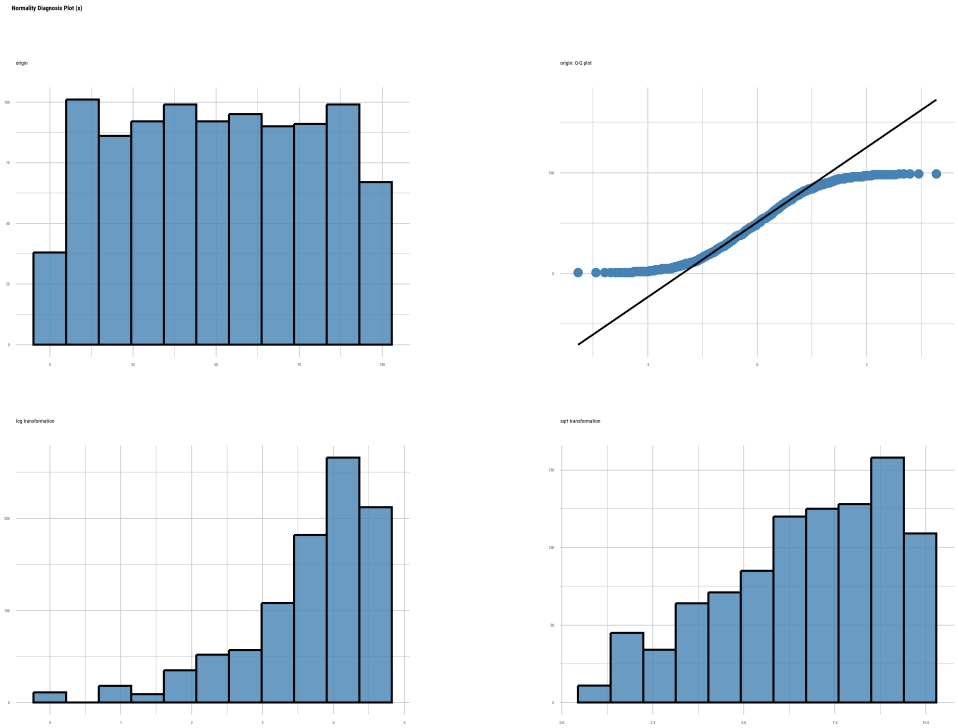
hours_per_week

statistic	p_value	remark
0.9526	6.1189e-17	No sample

Table 6: Shapiro-Wilk normality test

type	skewness	kurtosis
original	-0.0110	1.7995
log transformation	-1.5519	5.3589
sqrt transformation	-0.5591	2.3526

Table 6: skewness and kurtosis



Bivariate Analysis

Compare Numerical Variables

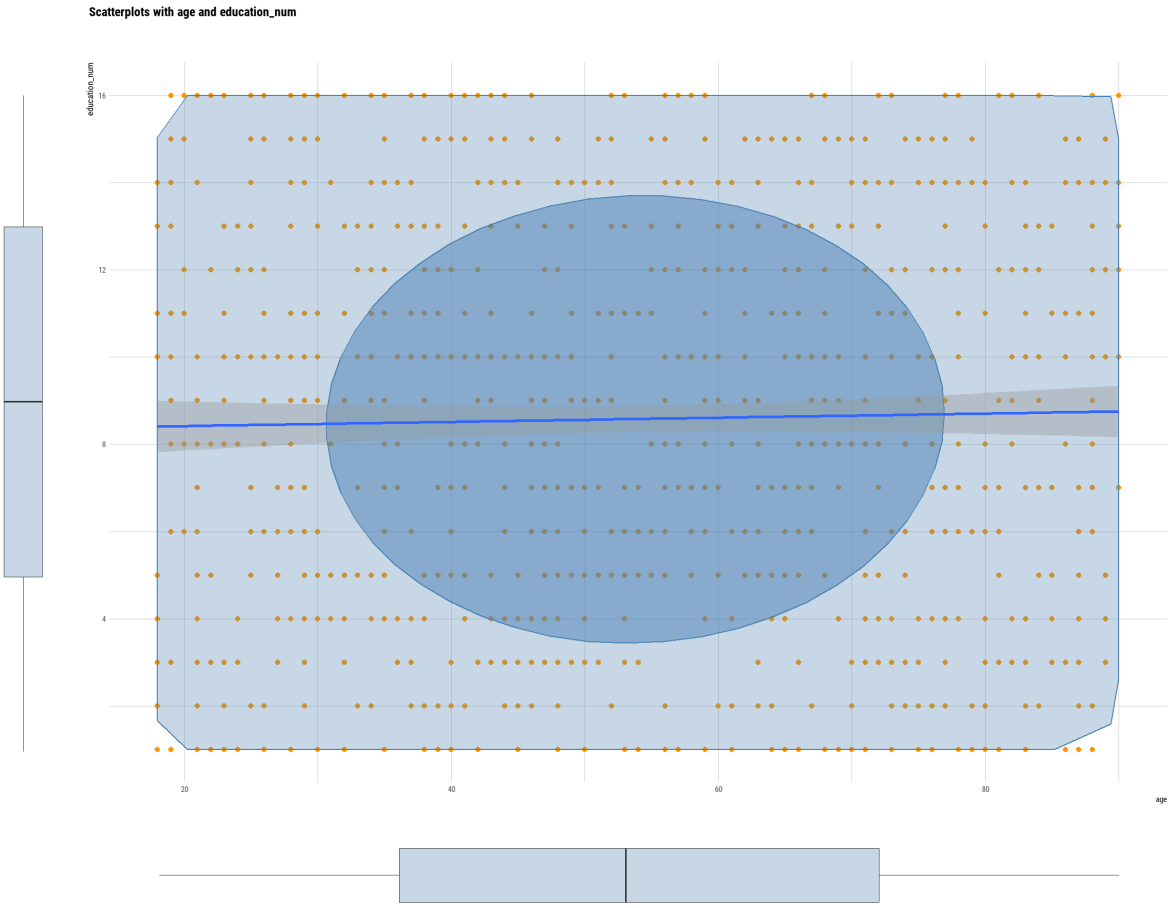
first variable	second variable	correlation coefficient
age	education_num	0.02160
age	capital_gain	-0.04186
age	capital_loss	0.03059
age	hours_per_week	0.00815
education_num	capital_gain	-0.00604
education_num	capital_loss	0.05562
education_num	hours_per_week	-0.02760
capital_gain	capital_loss	-0.07992
capital_gain	hours_per_week	-0.00760
capital_loss	hours_per_week	-0.02451

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	0.0004667	-0.0005877	20.99848	0.4426256	0.5060197	1

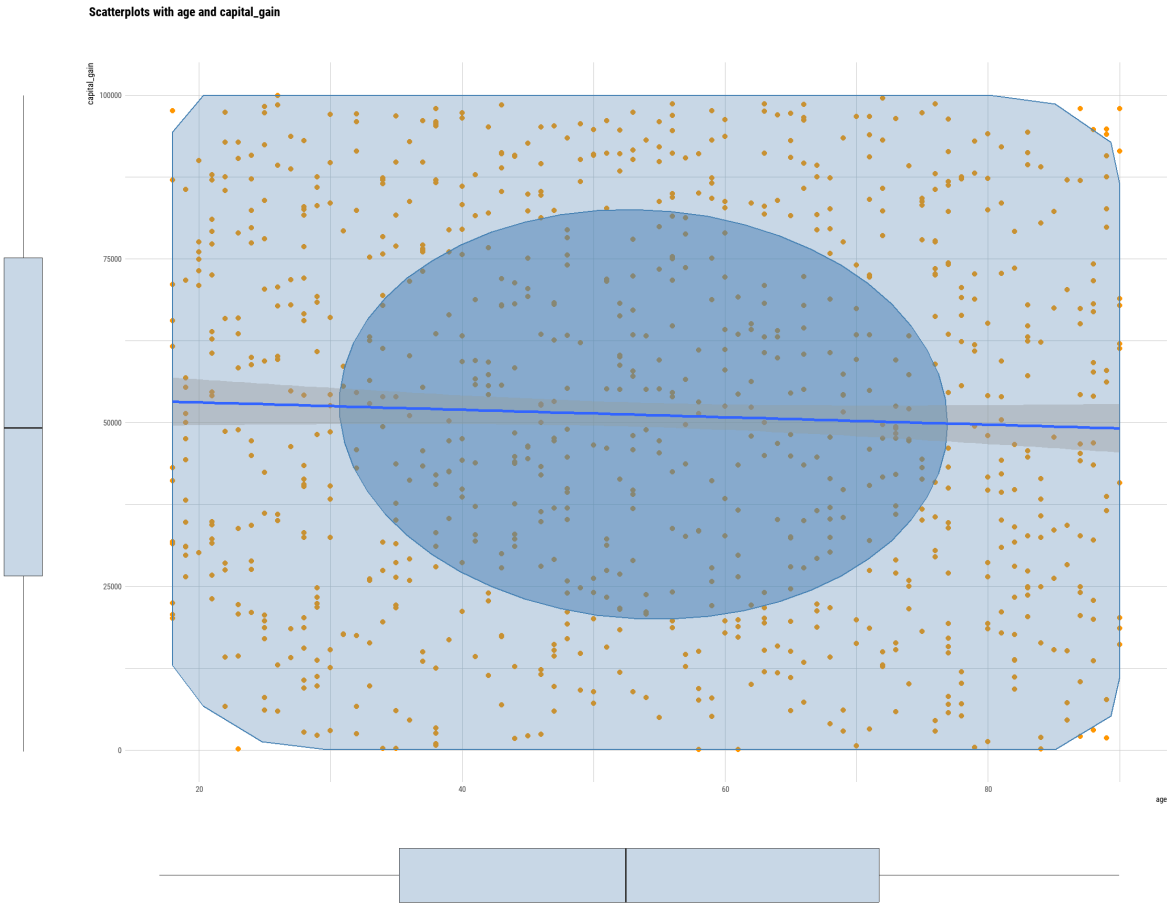
Table 7: Summary of linear model



‘age’ vs ‘capital_gain’

first variable	second variable	r.squared	adj.r.squared	sigma	statistic	p.value	df
age	capital_gain	0.0017526	0.0006483	20.90113	1.587131	0.2080617	1

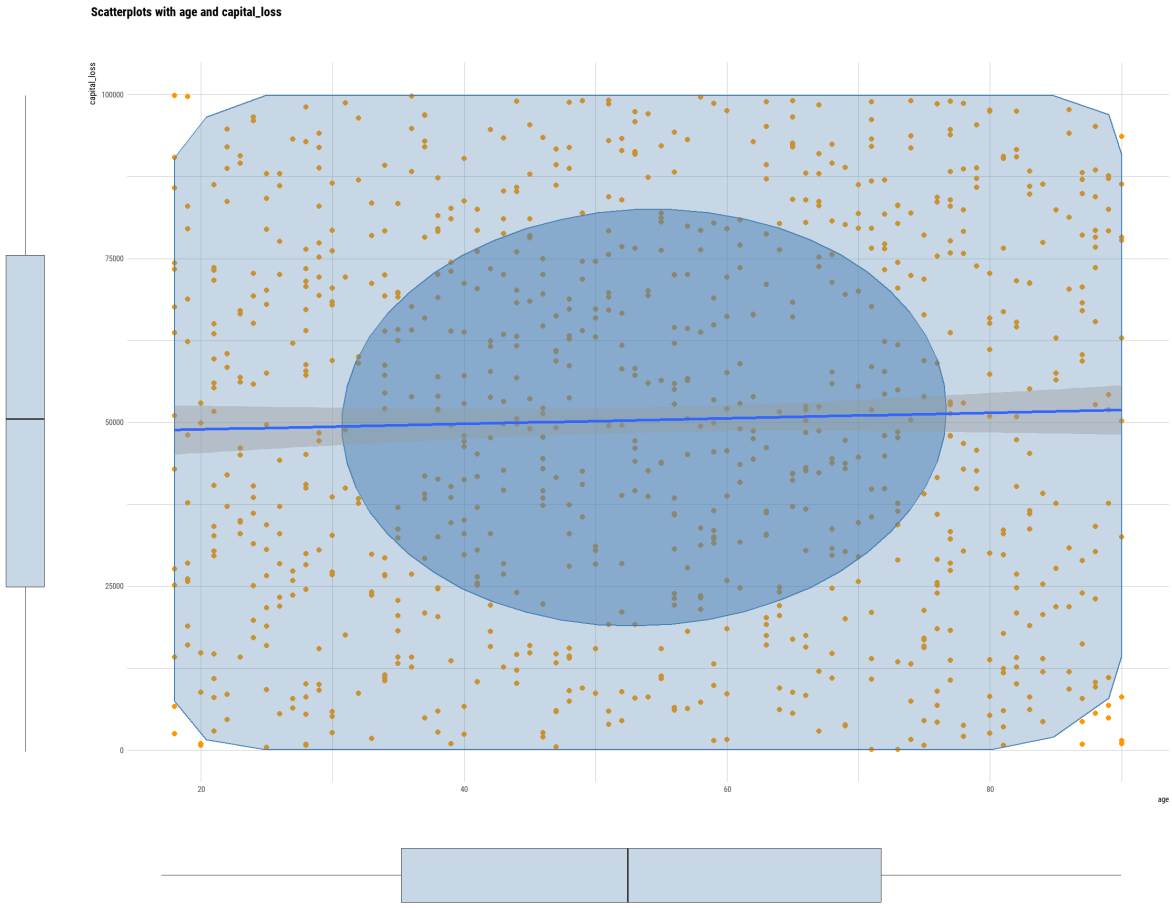
Table 7: Summary of linear model



‘age’ vs ‘capital_loss’

first variable	second variable	r.squared	adj.r.squared	sigma	statistic	p.value	df
age	capital_loss	0.000936	-0.0001704	20.9201	0.846013	0.3579283	1

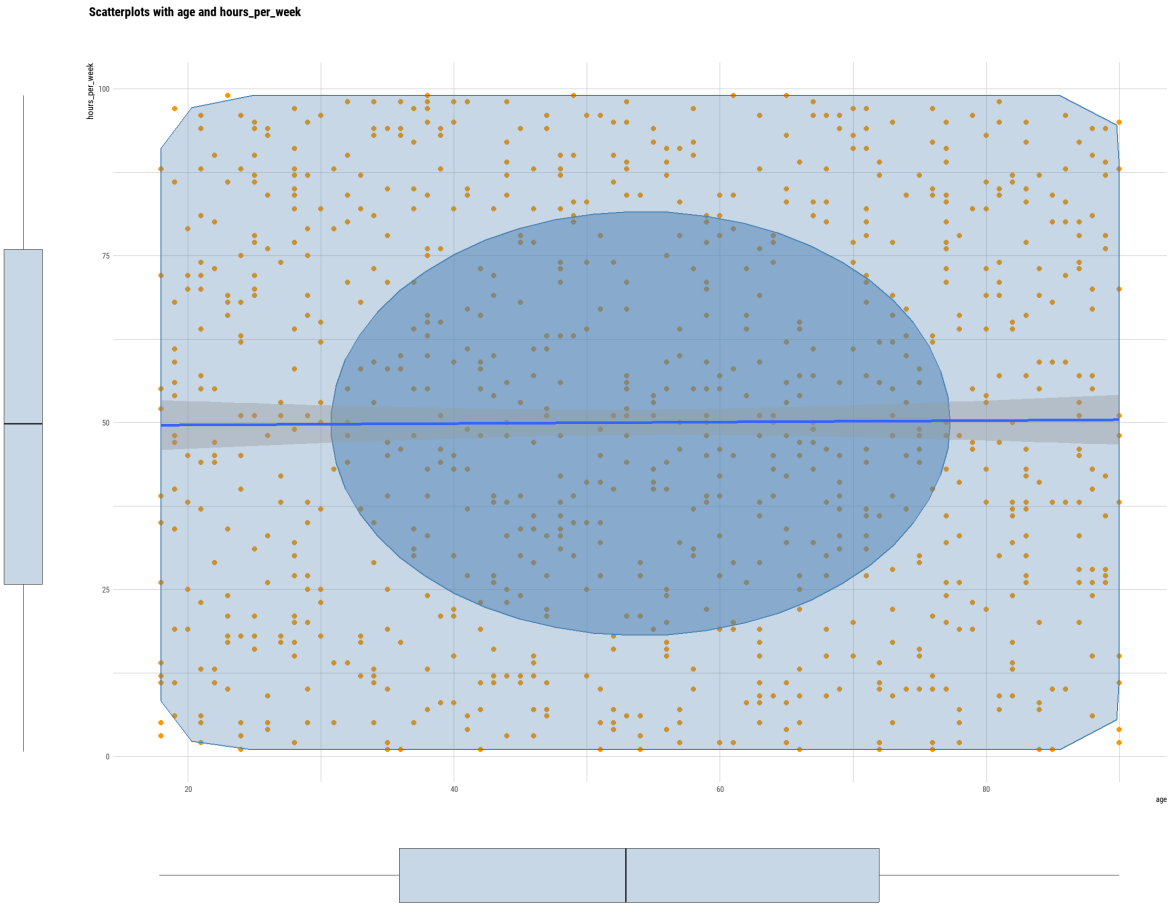
Table 7: Summary of linear model



‘age’ vs ‘hours_per_week’

first variable	second variable	r.squared	adj.r.squared	sigma	statistic	p.value	df
age	hours_per_week	6.64e-05	-0.0010471	21.11312	0.0596411	0.8071198	1

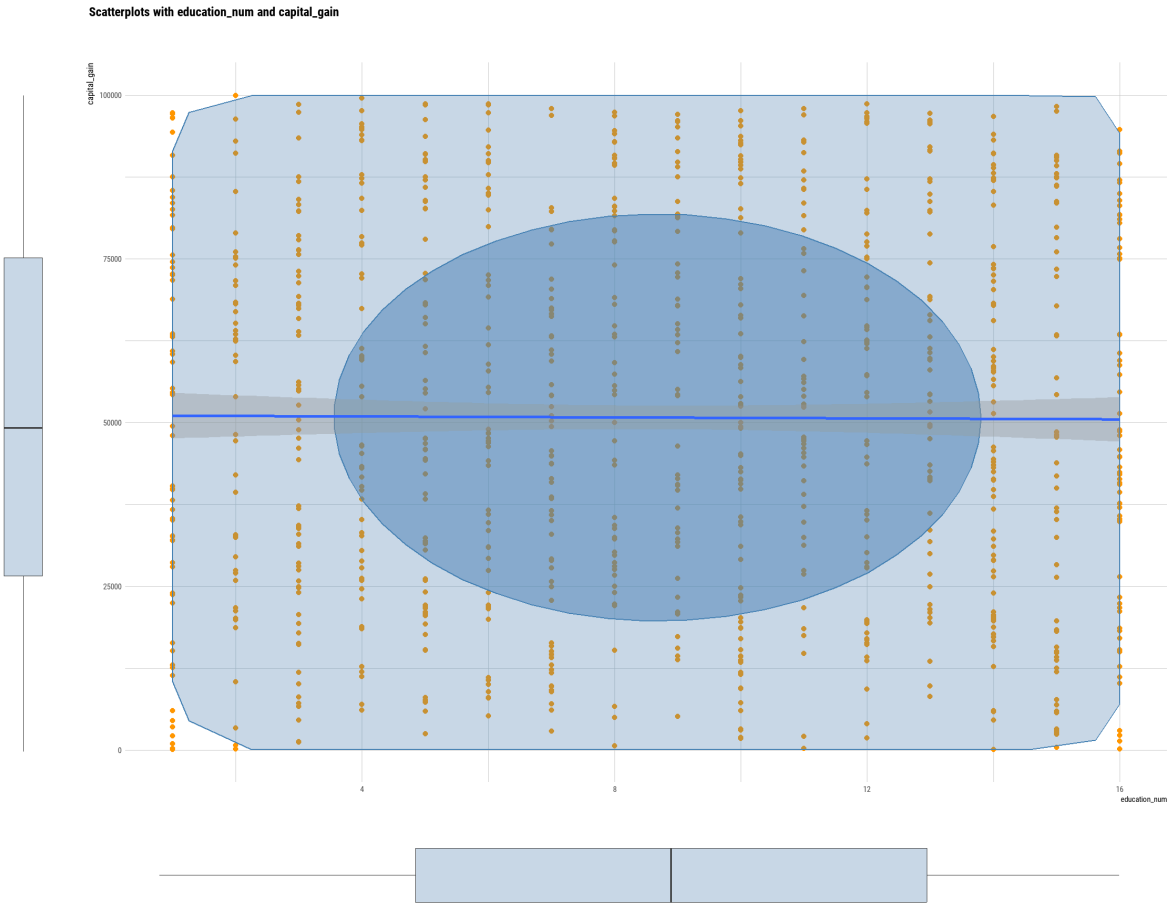
Table 7: Summary of linear model



‘education_num’ vs ‘capital_gain’

first variable	second variable	r.squared	adj.r.squared	sigma	statistic	p.value	df
education_num	capital_gain	3.65e-05	-0.0010183	4.642315	0.0345829	0.8525122	1

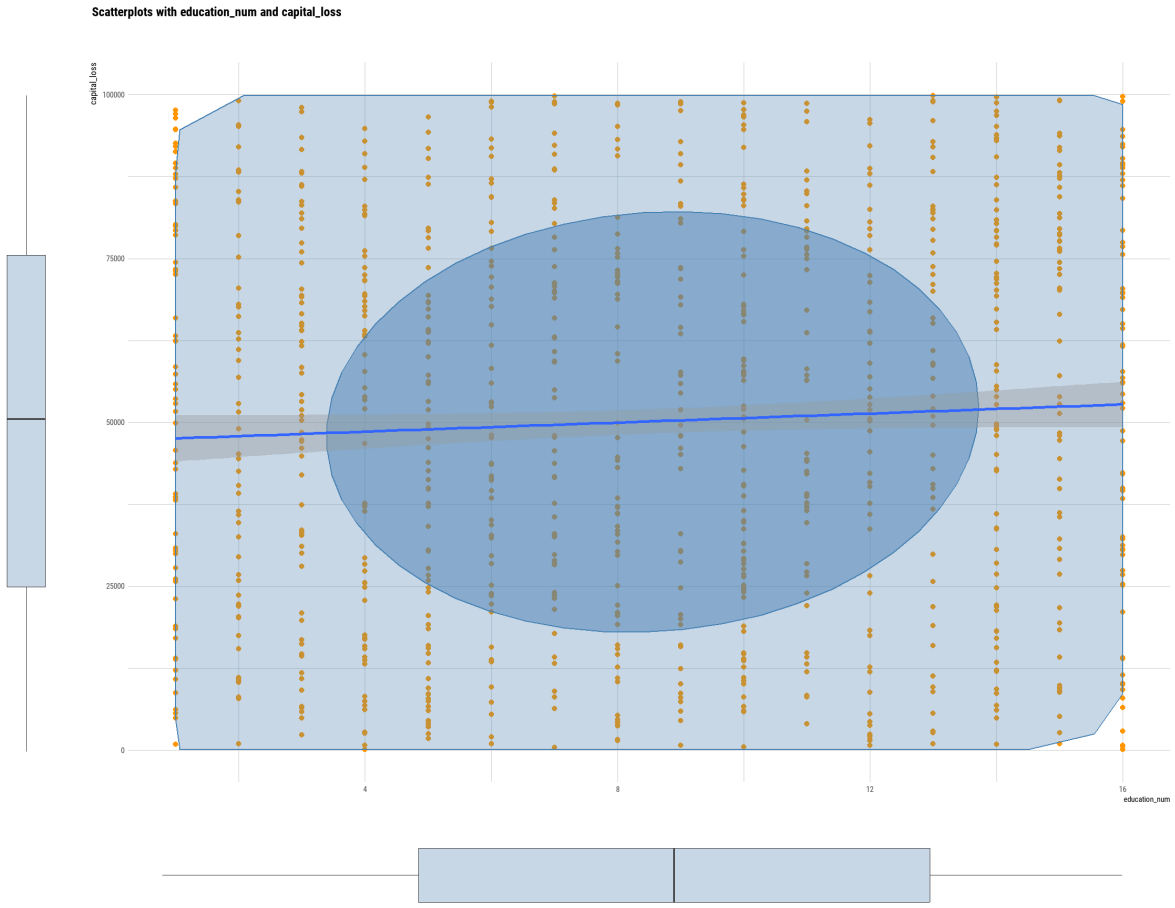
Table 7: Summary of linear model



‘education_num’ vs ‘capital_loss’

first variable	second variable	r.squared	adj.r.squared	sigma	statistic	p.value	df
education_num	capital_loss	0.0030941	0.0020425	4.668669	2.942314	0.0866138	1

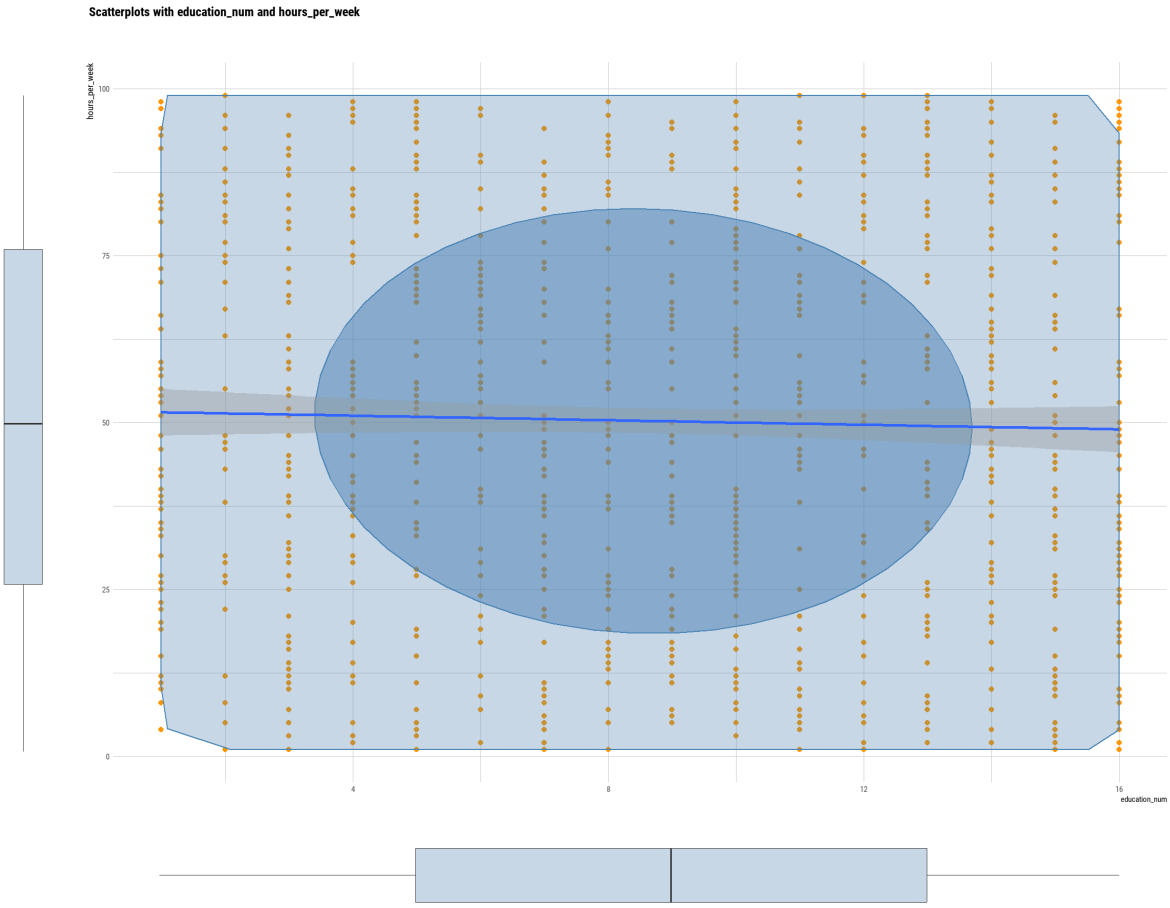
Table 7: Summary of linear model



‘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	0.0007618	-0.0002923	4.657871	0.7227186	0.3954683	1

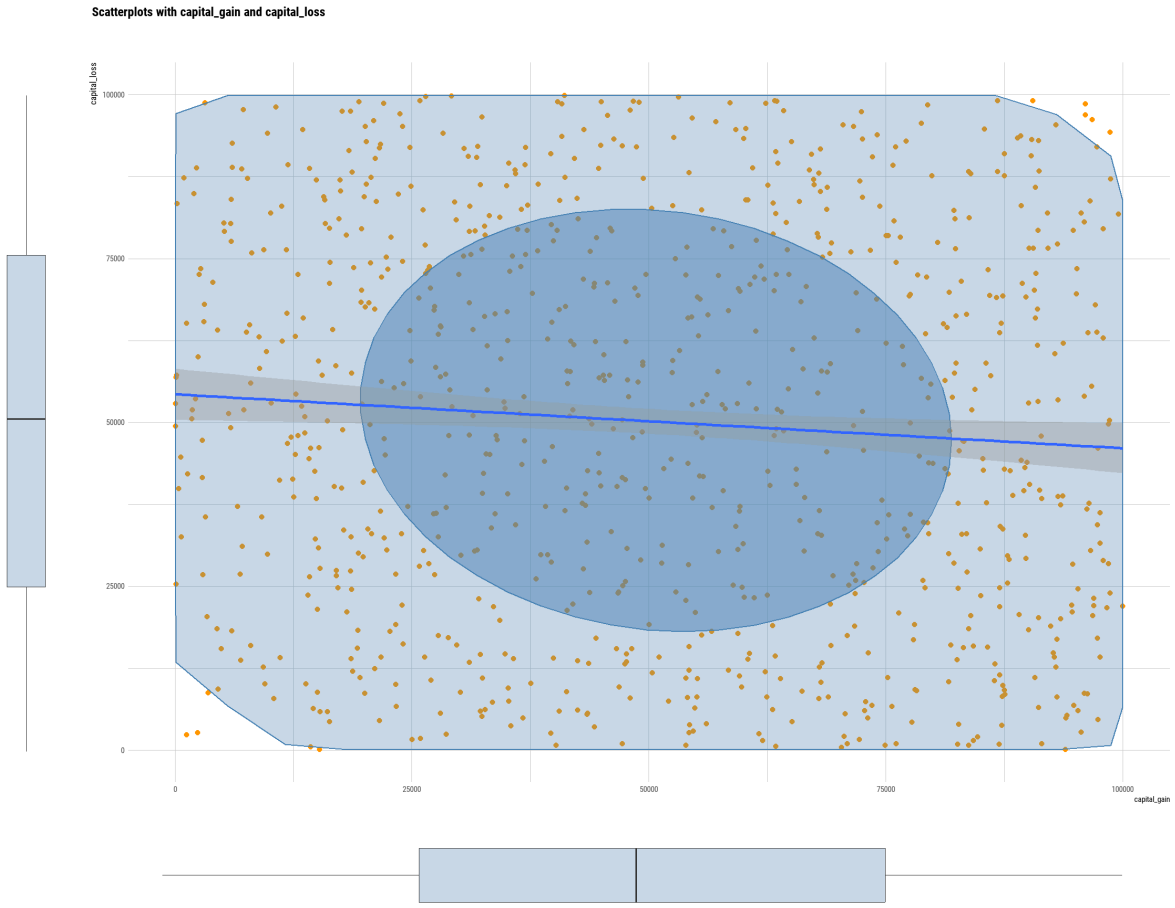
Table 7: Summary of linear model



‘capital_gain’ vs ‘capital_loss’

first variable	second variable	r.squared	adj.r.squared	sigma	statistic	p.value	df
capital_gain	capital_loss	0.0063879	0.0052876	28139.67	5.805372	0.0161765	1

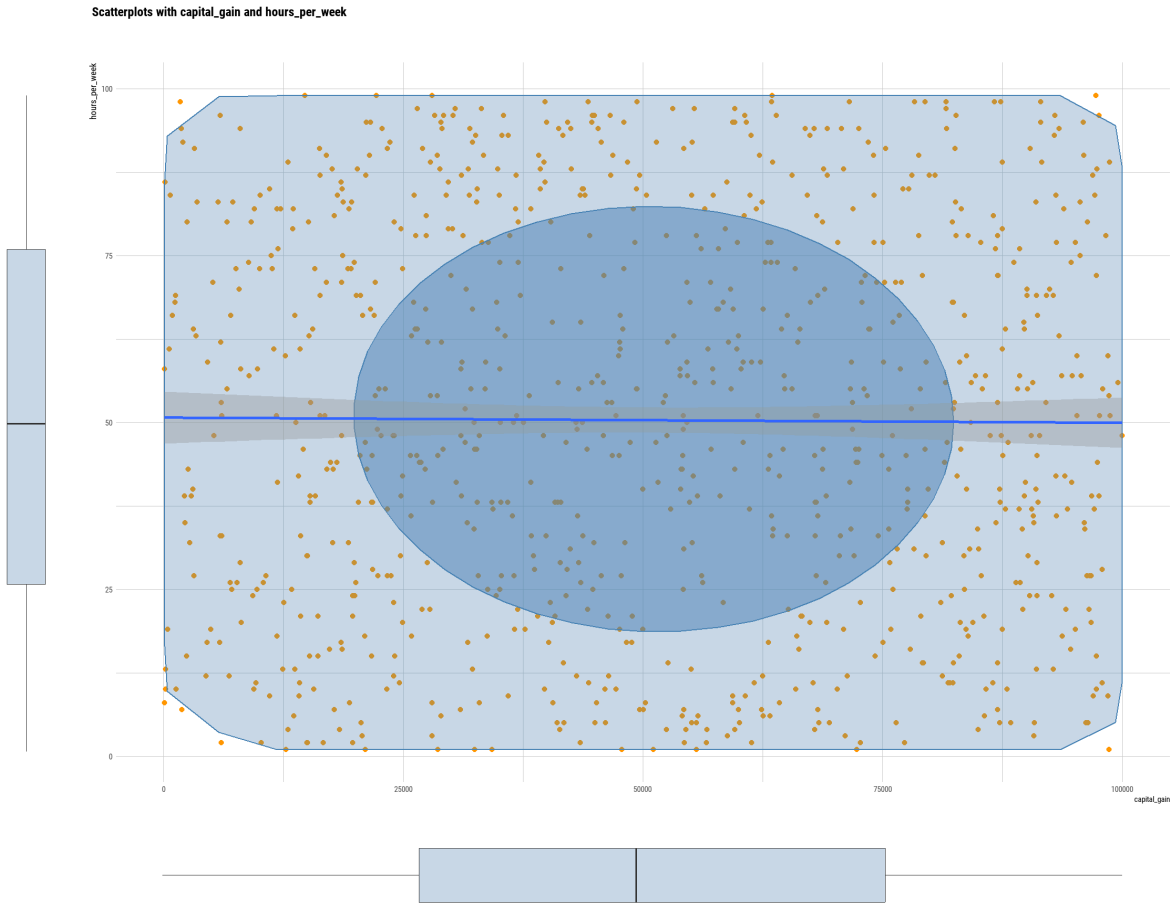
Table 7: Summary of linear model



‘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	5.77e-05	-0.0010558	28281.95	0.0518451	0.8199351	1

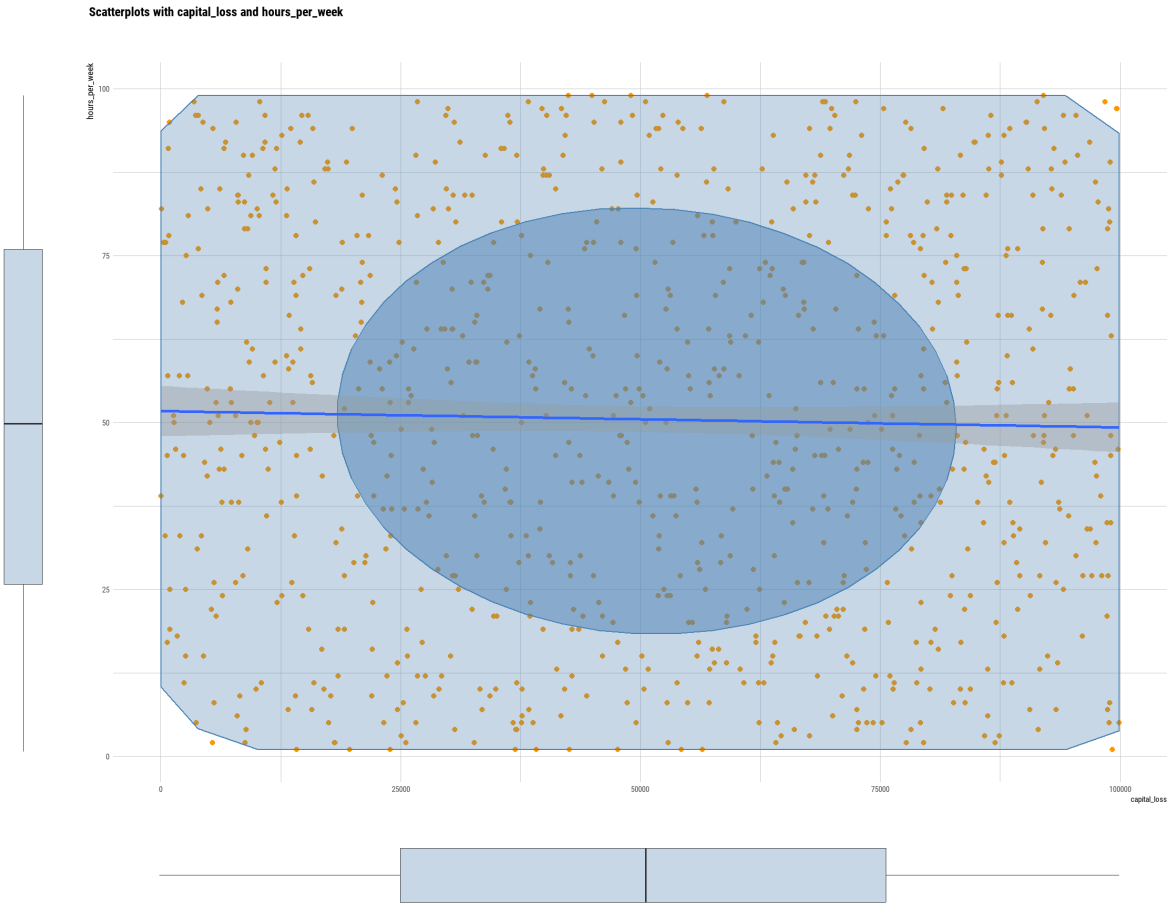
Table 7: Summary of linear model



‘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	0.0006009	-0.0005108	29124.93	0.5404921	0.4624206	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	0.022	-0.042	0.031	0.008
education_num	0.022	NA	-0.006	0.056	-0.028
capital_gain	-0.042	-0.006	NA	-0.080	-0.008
capital_loss	0.031	0.056	-0.080	NA	-0.025
hours_per_week	0.008	-0.028	-0.008	-0.025	NA

Table 8: Matrix table of correlation coefficient

Correlation Plot