MSCA31010: Linear & Non-Linear Models

Winter 2022 Assignment 4

The **Homeowner_Claim_History.xlsx** contains the claim history of 27,513 homeowner policies. The following table describes the eleven columns in the HOCLAIMDATA sheet.

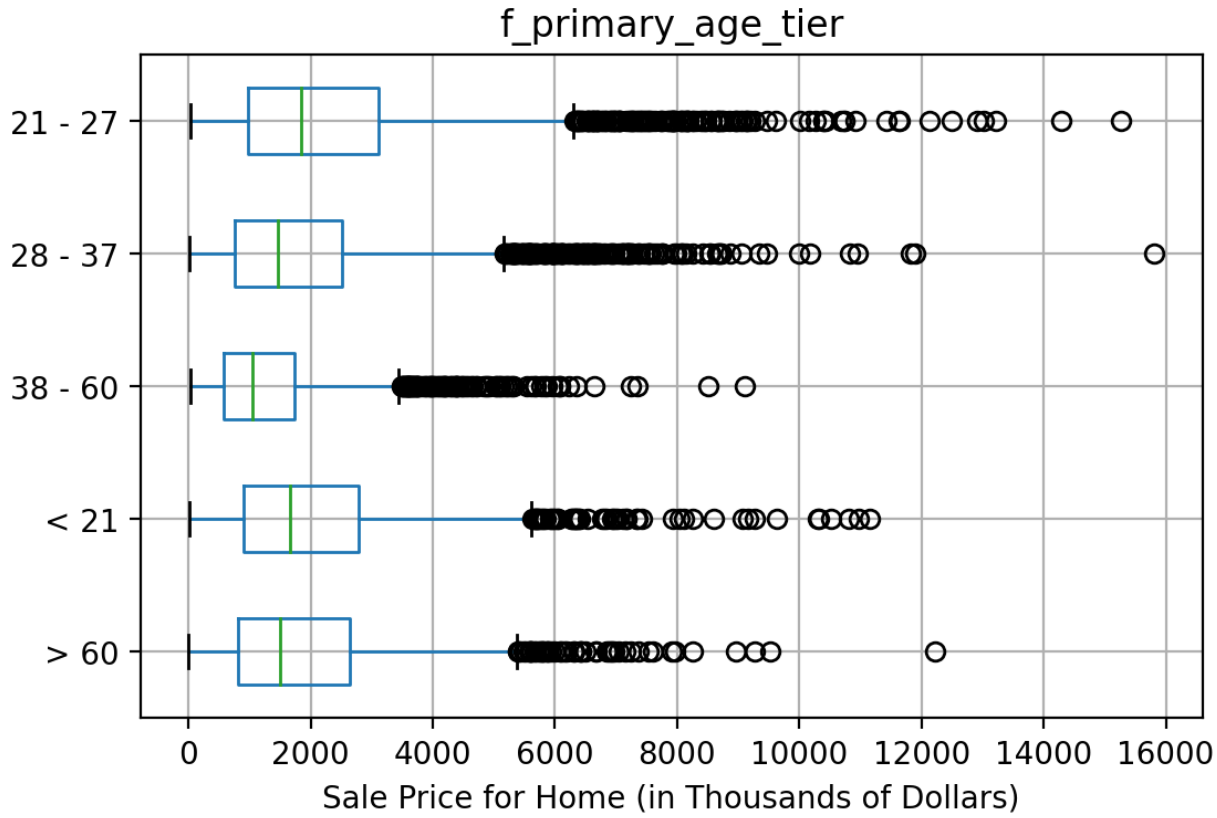
Name	Description	Categories
policy	Policy Identifier	
exposure	Duration a Policy is Exposed to Risk Measured in Portion of a Year	
num_claims	Number of Claims in a Year	
amt_claims	Total Claim Amount in a Year	
f_primary_age_tier	Age Tier of Primary Insured	< 21, 21 - 27, 28 - 37, 38 - 60, > 60
f_primary_gender	Gender of Primary Insured	Female, Male
f_marital	Marital Status of Primary Insured	Not Married, Married, Un-Married
f_residence_location	Location of Residence Property	Urban, Suburban, Rural
f_fire_alarm_type	Fire Alarm Type	None, Standalone, Alarm Service
f_mile_fire_station	Distance to Nearest Fire Station	< 1 mile, 1 - 5 miles, 6 - 10 miles, > 10 miles
f_aoi_tier	Amount of Insurance Tier	< 100K, 100K - 350K, 351K - 600K, 601K - 1M, > 1M

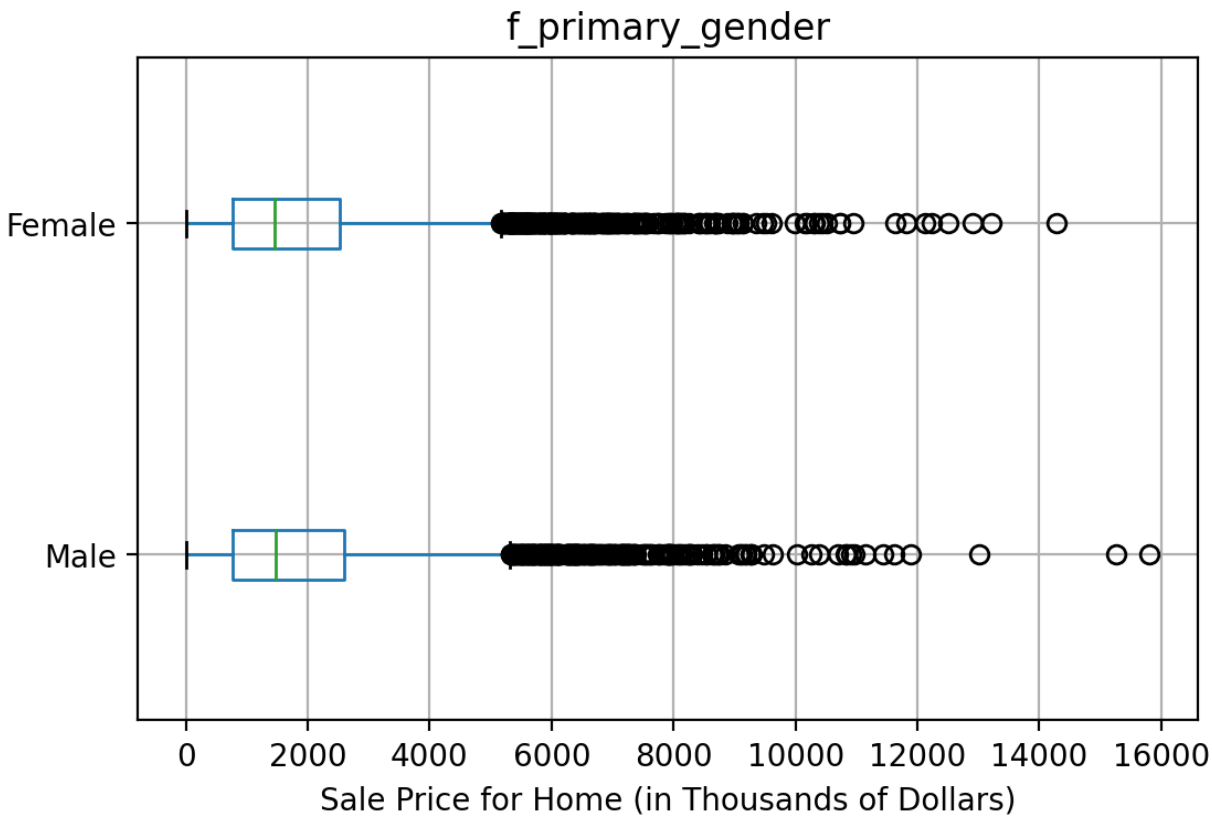
In insurance ratemaking, the ratio of Total Claim Amount in a Year divided by the Number of Claims in a Year is called the Severity. In other words, Severity is the average dollar amount per claim. If a policy does not file any claims in a year, then its Severity is missing.

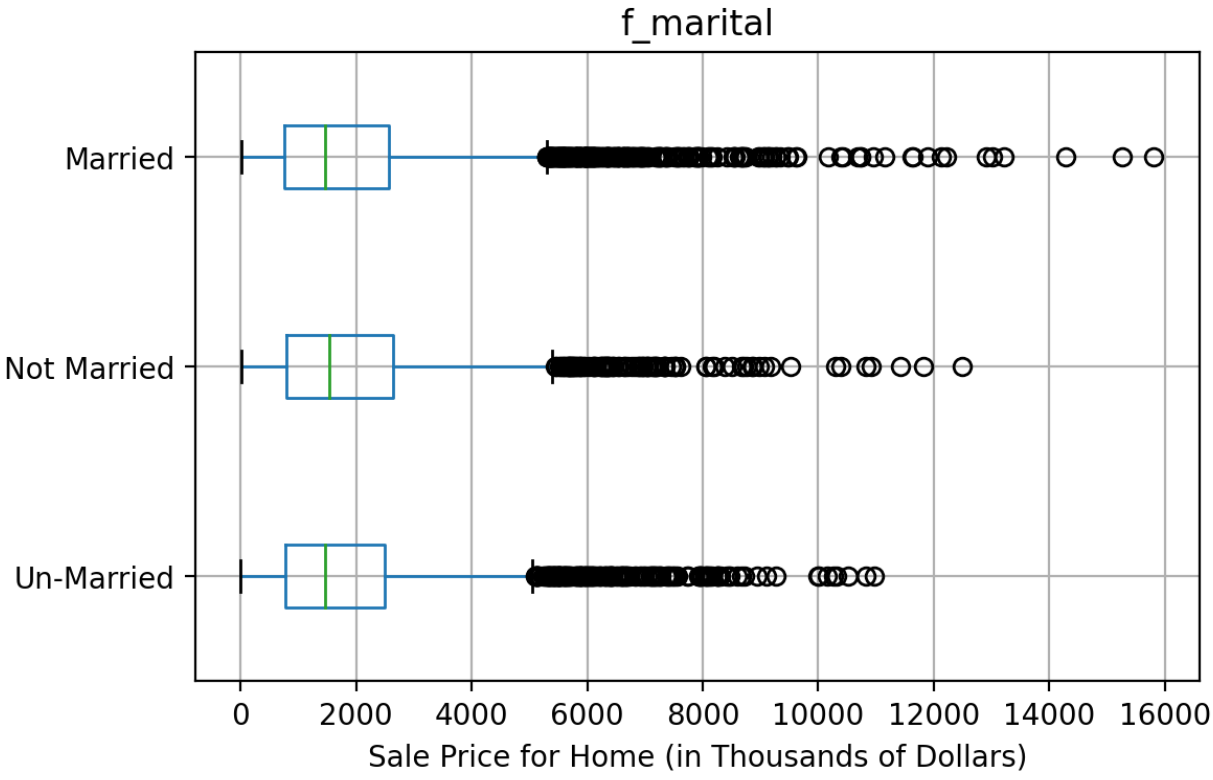
Unless otherwise stated, please provide all numeric answers rounded to the seventh decimal place.

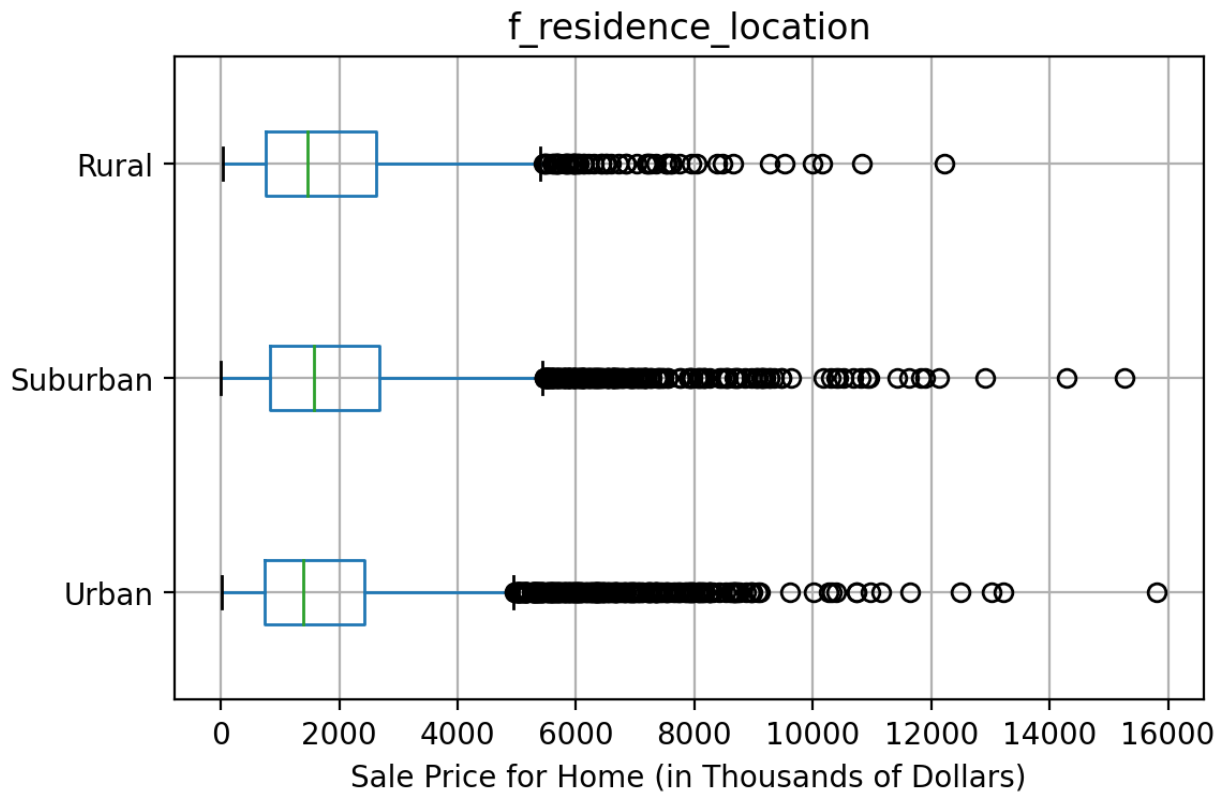
Question 1 (50 points)

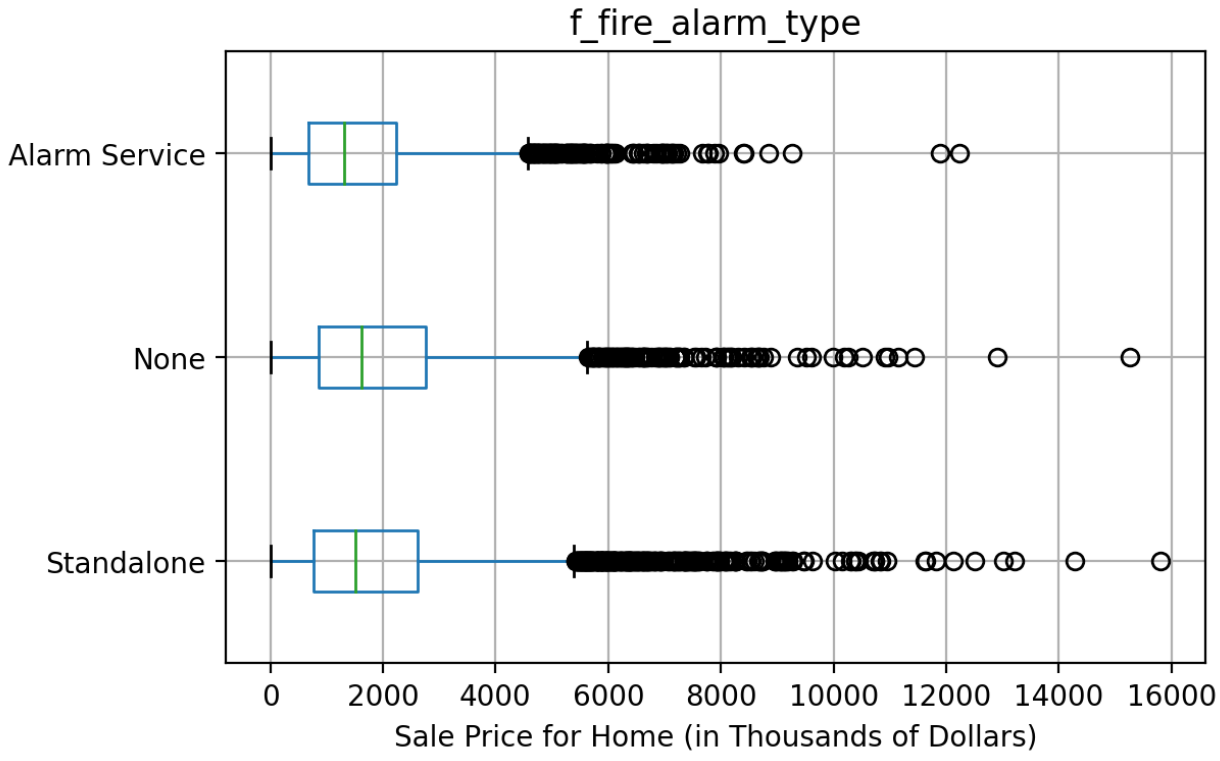
(a) (10 points) Generate horizontal boxplots of Total Claim Amount in a Year grouped by each of the seven categorical predictors *f_primary_age_tier*, *f_primary_gender*, *f_marital*, *f_residence_location*, *f_fire_alarm_type*, *f_mile_fire_station*, and *f_aoi_tier*.

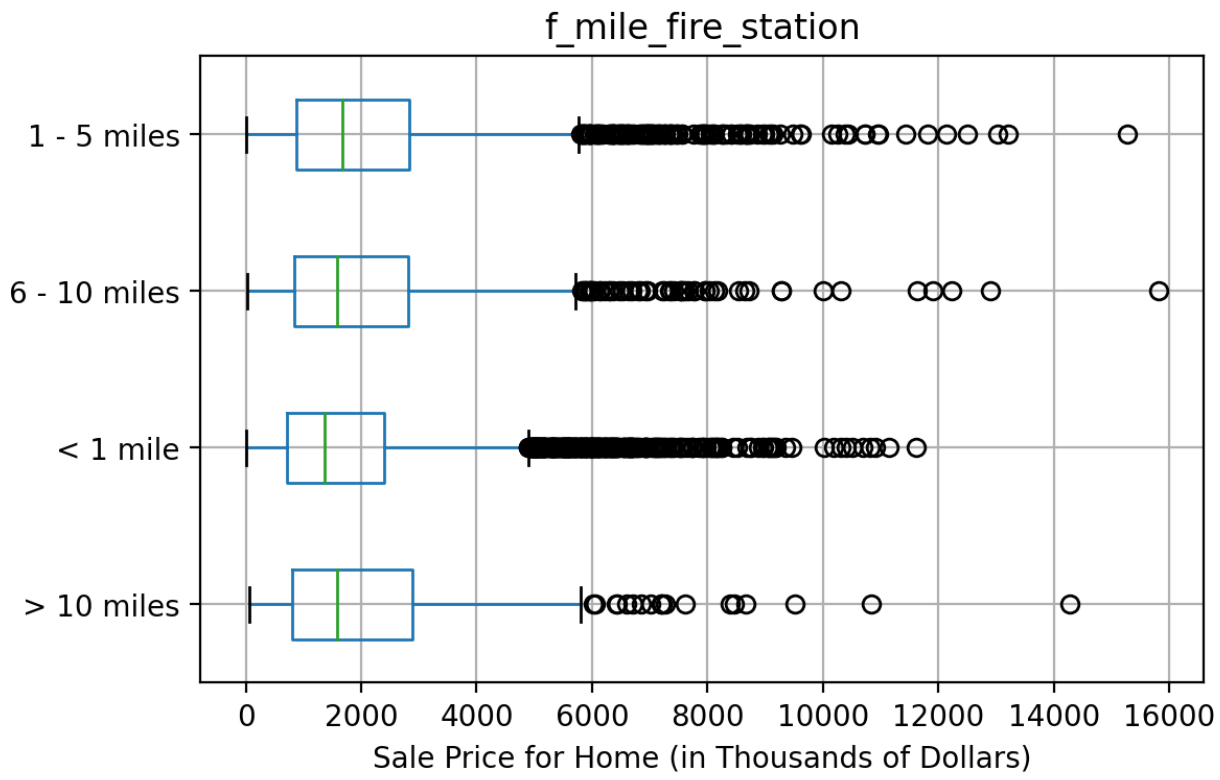


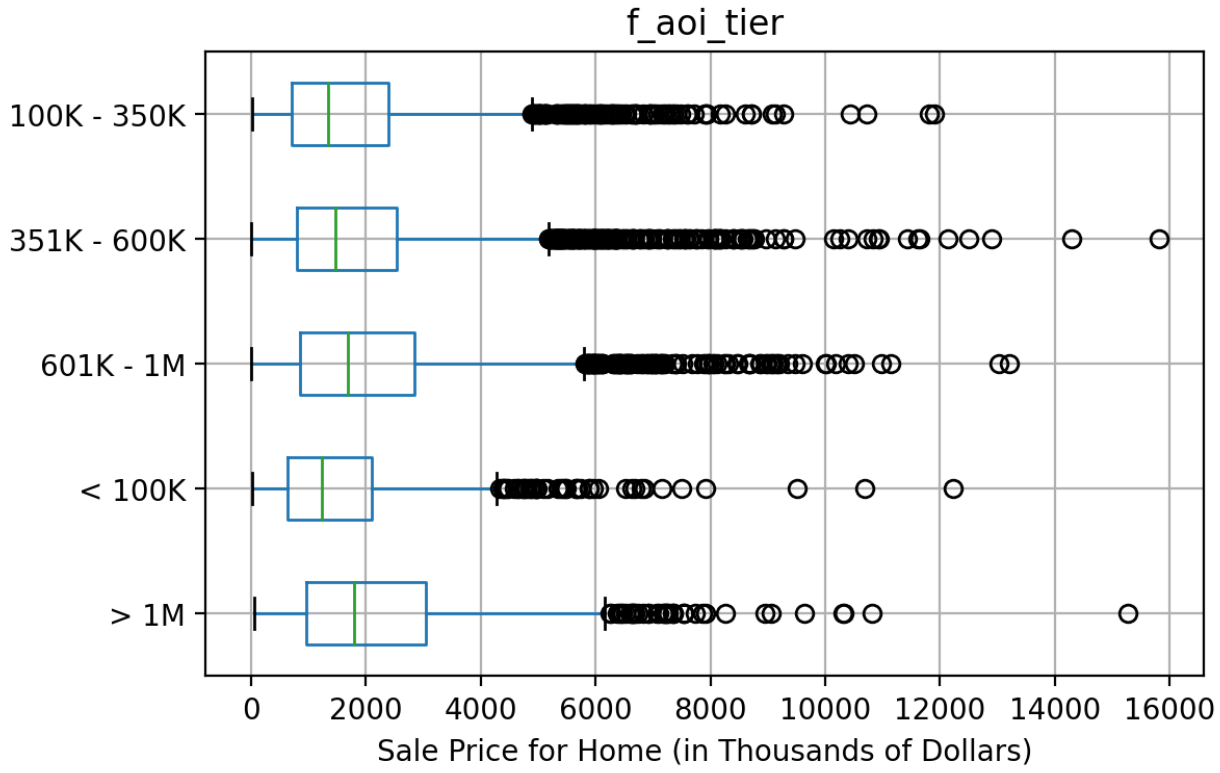

f_primary_gender
Female
Male
0
2000
4000
6000
8000
10000
12000
14000
16000
Sale Price for Home (in Thousands of Dollars)


f_marital
Married
Not Married
Un-Married
0
2000
4000
6000
8000
10000
12000
14000
16000
Sale Price for Home (in Thousands of Dollars)


f_residence_location
Rural
Suburban
Urban
0
2000
4000
6000
8000
10000
12000
14000
16000
Sale Price for Home (in Thousands of Dollars)


f_fire_alarm_type
Alarm Service
None
Standalone
0
2000
4000
6000
8000
10000
12000
14000
16000
Sale Price for Home (in Thousands of Dollars)


f_mile_fire_station
1 - 5 miles
6 - 10 miles
< 1 mile
> 10 miles
0
2000
4000
6000
8000
10000
12000
14000
16000
Sale Price for Home (in Thousands of Dollars)


f_aoi_tier
100K - 350K
351K - 600K
601K - 1M
< 100K
> 1M
0
2000
4000
6000
8000
10000
12000
14000
16000
Sale Price for Home (in Thousands of Dollars)

(b) (10 points) For analyses, Severity will follow a Gamma distribution. Train a Gamma model with the logarithm link function. The target variable is Severity (use only positive and non-missing values for analyses). The predictors are the seven categorical predictors. The model will include the Intercept term. Enter predictors into the model using the Forward Selection method. The entry threshold is 0.05. What is the estimate for the Shape parameter?

======= Step Detail =======

Step = 1

Step Statistics:

	Predictor	Type	ModelDF	ModelLLK	DevChiSq	DevDF	DevSig
0	f_primary_age_tier	categorical	5	-99,738.72	404.3866	4	3.136693e-86
6	f_aoi_tier	categorical	5	-99,886.54	108.747	4	1.346463e-22
4	f_fire_alarm_type	categorical	3	-99,911.79	58.24039	2	2.25559e-13
5	f_mile_fire_station	categorical	4	-99,911.68	58.46986	3	1.247624e-12
3	f_residence_location	categorical	3	-99,921.65	38.52957	2	4.299427e-09
2	f_marital	categorical	3	-99,938.05	5.720389	2	0.05725763
1	f_primary_gender	categorical	2	-99,940.54	0.7429267	1	0.3887249

Enter predictor = f_primary_age_tier

Minimum P-Value = 3.1366925636835666e-86

======= Step Detail =======

Step = 2

Step Statistics:

	Predictor	Type	ModelDF	ModelLLK	DevChiSq	DevDF	DevSig
5	f_aoi_tier	categorical	9	-99,679.7	118.0392	4	1.400917e-24
3	f_fire_alarm_type	categorical	7	-99,704.02	69.39581	2	8.528872e-16
4	f_mile_fire_station	categorical	8	-99,708.23	60.97377	3	3.640657e-13
2	f_residence_location	categorical	7	-99,718.89	39.64993	2	2.455426e-09
1	f_marital	categorical	7	-99,736.23	4.973669	2	0.08317283
0	f_primary_gender	categorical	6	-99,738.44	0.5629141	1	0.4530885

Enter predictor = f_aoi_tier

Minimum P-Value = 1.4009169837758559e-24

======= Step Detail =======

Step = 3

Step Statistics:

	Predictor	Type	ModelDF	ModelLLK	DevChiSq	DevDF	DevSig
3	f_fire_alarm_type	categorical	11	-99,643.66	72.07999	2	2.228579e-16
4	f_mile_fire_station	categorical	12	-99,648.34	62.72641	3	1.536613e-13
2	f_residence_location	categorical	11	-99,659.33	40.74656	2	1.419048e-09
1	f_marital	categorical	11	-99,677.16	5.079599	2	0.07888221
0	f_primary_gender	categorical	10	-99,679.49	0.4105454	1	0.5216928

Enter predictor = f_fire_alarm_type

Minimum P-Value = 2.2285791437852406e-16

======= Step Detail =======

Step = 4

Step Statistics:

	Predictor	Type	ModelDF	ModelLLK	DevChiSq	DevDF	DevSig
3	f_mile_fire_station	categorical	14	-99,610.62	66.06938	3	2.962017e-14
2	f_residence_location	categorical	13	-99,622.45	42.42847	2	6.12034e-10
1	f_marital	categorical	13	-99,641.17	4.974828	2	0.08312465
0	f_primary_gender	categorical	12	-99,643.46	0.4034187	1	0.525329

Enter predictor = f_mile_fire_station

Minimum P-Value = 2.962016608253086e-14

======= Step Detail =======

Step = 5

Step Statistics:

	Predictor	Type	ModelDF	ModelLLK	DevChiSq	DevDF	DevSig
2	f_residence_location	categorical	16	-99,585.69	49.87939	2	1.475124e-11
1	f_marital	categorical	16	-99,607.96	5.335108	2	0.06942182
0	f_primary_gender	categorical	15	-99,610.36	0.5343199	1	0.4647963

Enter predictor = f_residence_location

Minimum P-Value = 1.4751239916365028e-11

The estimate for the Shape parameter is 2.1725703

(c) (10 points) Provide the Step Summary table. The table should contain (1) Step Number, (2) Model Degrees of Freedom, (3) Model Log-Likelihood, (4) Deviance Chi-Squares, (5) Deviance Degrees of Freedom, and (6) Deviance Significance. Show the Significance in .E7 scientific notation.

	Predictor	Type	ModelDF	ModelLLK	DevChiSq	DevDF	DevSig
0	Intercept		1	-99,940.91	NaN	NaN	NaN
1	f_primary_age_tier	categorical	5	-99,738.72	404.3866	4.0	3.136693e-86
2	f_aoi_tier	categorical	9	-99,679.7	118.0392	4.0	1.400917e-24
3	f_fire_alarm_type	categorical	11	-99,643.66	72.07999	2.0	2.228579e-16
4	f_mile_fire_station	categorical	14	-99,610.62	66.06938	3.0	2.962017e-14
5	f_residence_location	categorical	16	-99,585.69	49.87939	2.0	1.475124e-11

(d) (10 points) Assess the final model goodness-of-fit using (1) Root Mean Squared Error, (2) Relative Error, (3) Mean Absolute Proportion Error, and (4) Pearson Correlation. What are the values of these metrics?

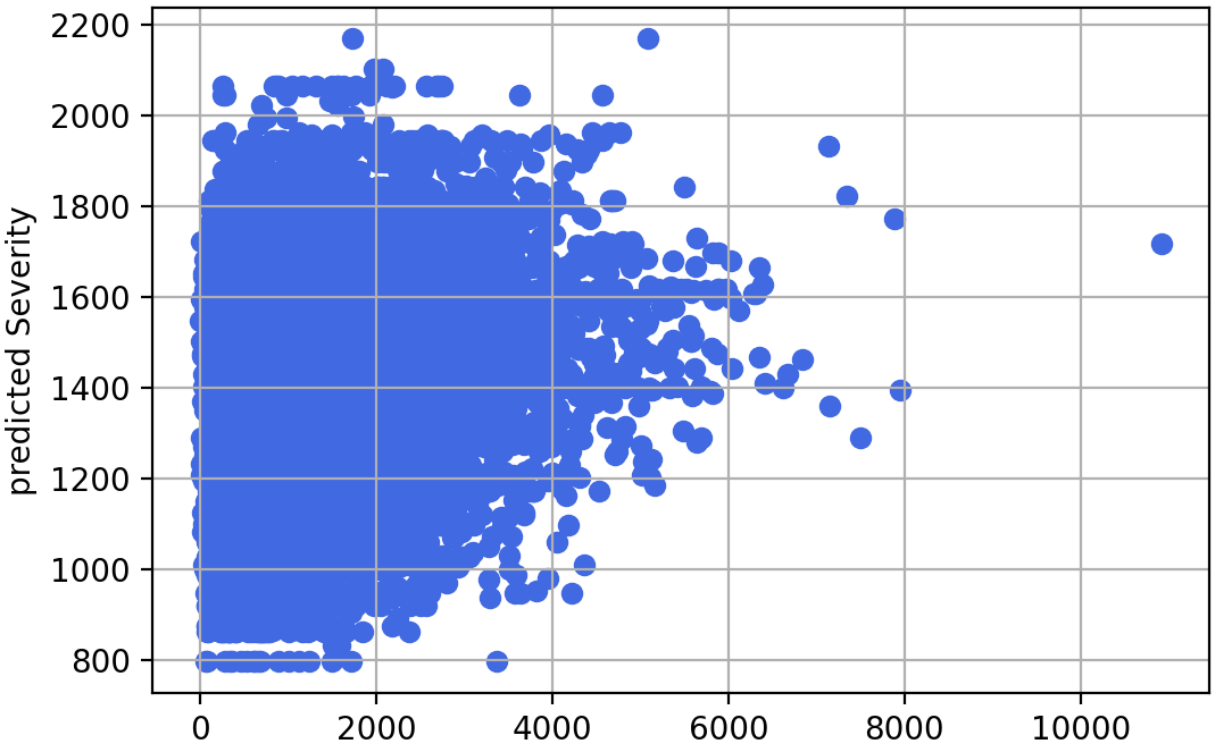
Root Mean Squared Error = 942.2956575

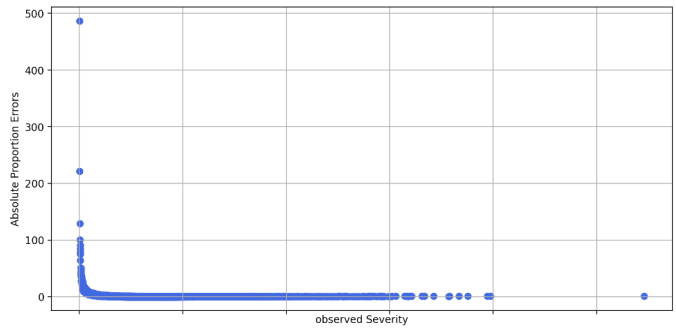
Relative Error = 0.9468733

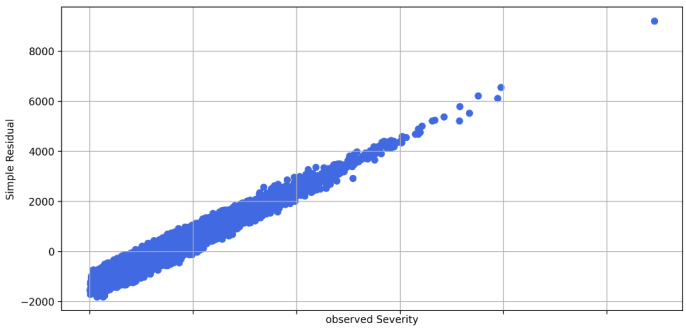
Mean Absolute Proportion Error = 1.2073721

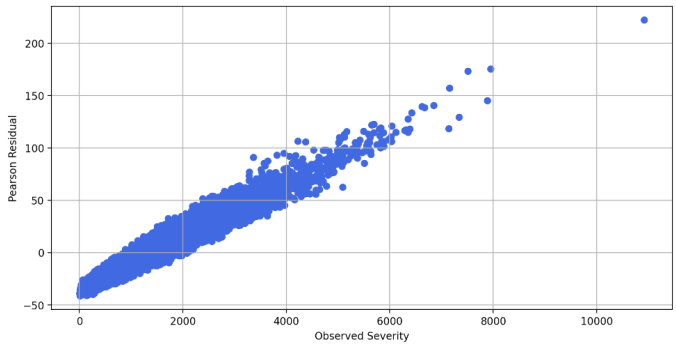
Pearson Correlation = 0.2305204

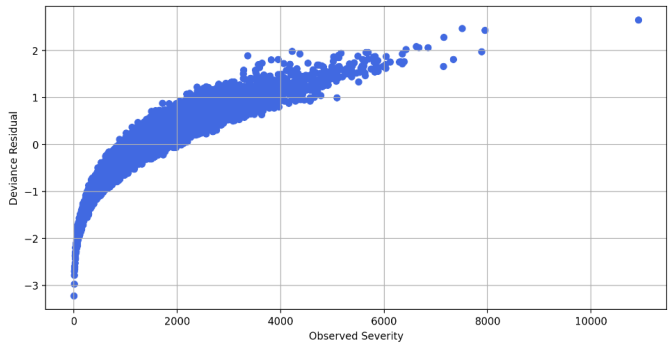
(e) (10 points) Identify any poorly predicted observations. First, plot the predicted versus the observed Severity. Second, together in a single chart frame, plot the Simple Residuals, the Pearson Residuals, the Deviance Residuals, and the Absolute Proportion Errors versus the observed Severity. Label the axes of these two charts accordingly. To receive full credits, generate your charts with proper dimensions (e.g., length and width) and resolution (e.g., dpi).











There is one poorly predicted observation.

Question 2 (50 points)

(a) (20 points). Train a Multi-Layer Perceptron neural network. The target variable is Severity (use only positive and non-missing values for analyses). The predictors are the seven categorical predictors. Perform a naïve grid search to select the best network structure. For each Hyperbolic Tangent and Rectified Linear Unit activation function, try the number of layers from 1 to 10, the common number of neurons per layer from 1 to 5. Provide a table that shows your grid search results. The table should contain (1) the activation function type, (2) the number of layers, (3) the common number of neurons per layer, (4) the total number of neurons, and (5) the mean absolute proportion error.

	Activation Function	nLayer	nHiddenNeuron	Total Number of Neuron	MAPE	Elapsed Time
48	relu	10	4	(4, 4, 4, 4, 4, 4, 4, 4, 4, 4)	1.178096	6.196255
36	relu	8	2	(2, 2, 2, 2, 2, 2, 2, 2)	1.189122	2.951748
28	relu	6	4	(4, 4, 4, 4, 4, 4)	1.189456	5.460685
27	relu	6	3	(3, 3, 3, 3, 3, 3)	1.191035	2.881814
47	relu	10	3	(3, 3, 3, 3, 3, 3, 3, 3, 3, 3)	1.191088	2.162976
32	relu	7	3	(3, 3, 3, 3, 3, 3, 3)	1.193823	2.764959
19	relu	4	5	(5, 5, 5, 5)	1.194961	3.995578
29	relu	6	5	(5, 5, 5, 5, 5, 5)	1.19888	3.124536
38	relu	8	4	(4, 4, 4, 4, 4, 4, 4, 4)	1.202424	5.172626
2	relu	1	3	(3,)	1.203192	24.47467
21	relu	5	2	(2, 2, 2, 2, 2)	1.203441	2.579425

9	relu	2	5	(5, 5)	1.204006	8.384083
4	relu	1	5	(5,)	1.204435	19.51022
7	relu	2	3	(3, 3)	1.204836	4.078422
16	relu	4	2	(2, 2, 2, 2)	1.204994	18.0529
1	relu	1	2	(2,)	1.206532	29.48747
8	relu	2	4	(4, 4)	1.206546	6.623947
3	relu	1	4	(4,)	1.206566	8.533017
14	relu	3	5	(5, 5, 5)	1.206599	4.683151
13	relu	3	4	(4, 4, 4)	1.206792	4.343649
0	relu	1	1	(1,)	1.207094	7.349537
18	relu	4	4	(4, 4, 4, 4)	1.207922	4.84369
12	relu	3	3	(3, 3, 3)	1.209825	3.917867
39	relu	8	5	(5, 5, 5, 5, 5, 5, 5, 5)	1.212128	4.169229
11	relu	3	2	(2, 2, 2)	1.212258	3.333801
33	relu	7	4	(4, 4, 4, 4, 4, 4, 4)	1.213168	4.647314
34	relu	7	5	(5, 5, 5, 5, 5, 5, 5)	1.214071	3.606315
26	relu	6	2	(2, 2, 2, 2, 2, 2)	1.21812	3.210088
23	relu	5	4	(4, 4, 4, 4, 4)	1.222937	4.558918

44	relu	9	5	(5, 5, 5, 5, 5, 5, 5, 5, 5, 5)	1.262297	4.535128
30	relu	7	1	(1, 1, 1, 1, 1, 1, 1)	1.262493	2.506892
37	relu	8	3	(3, 3, 3, 3, 3, 3, 3, 3)	1.262716	2.14584
43	relu	9	4	(4, 4, 4, 4, 4, 4, 4, 4, 4)	1.262821	7.318351
17	relu	4	3	(3, 3, 3, 3)	1.26318	27.83101
25	relu	6	1	(1, 1, 1, 1, 1, 1)	1.26323	4.196953
49	relu	10	5	(5, 5, 5, 5, 5, 5, 5, 5, 5, 5)	1.26324	8.651152
40	relu	9	1	(1, 1, 1, 1, 1, 1, 1, 1, 1)	1.263317	19.23787
31	relu	7	2	(2, 2, 2, 2, 2, 2, 2)	1.263335	21.37269
10	relu	3	1	(1, 1, 1)	1.263385	8.753487
46	relu	10	2	(2, 2, 2, 2, 2, 2, 2, 2, 2, 2)	1.26342	27.17368
6	relu	2	2	(2, 2)	1.263462	6.904228
42	relu	9	3	(3, 3, 3, 3, 3, 3, 3, 3, 3)	1.26367	1.492412
45	relu	10	1	(1, 1, 1, 1, 1, 1, 1, 1, 1, 1)	1.263707	5.683859
15	relu	4	1	(1, 1, 1, 1)	1.26376	2.851327

(b) (10 points) Recommend the best network structure which yields the lowest Mean Absolute Proportion Error. In the case of ties, choose the network with a fewer total number of neurons.

Activation Function: relu

nLayer: 10

nHiddenNeuron:4

Total Number of Neuron:40

MAPE: 1.178096

Elapsed Time: 6.196255

(c) (10 points) Assess the final model goodness-of-fit using (1) Root Mean Squared Error, (2) Relative Error, (3) Mean Absolute Proportion Error, and (4) Pearson Correlation. What are the values of these metrics?

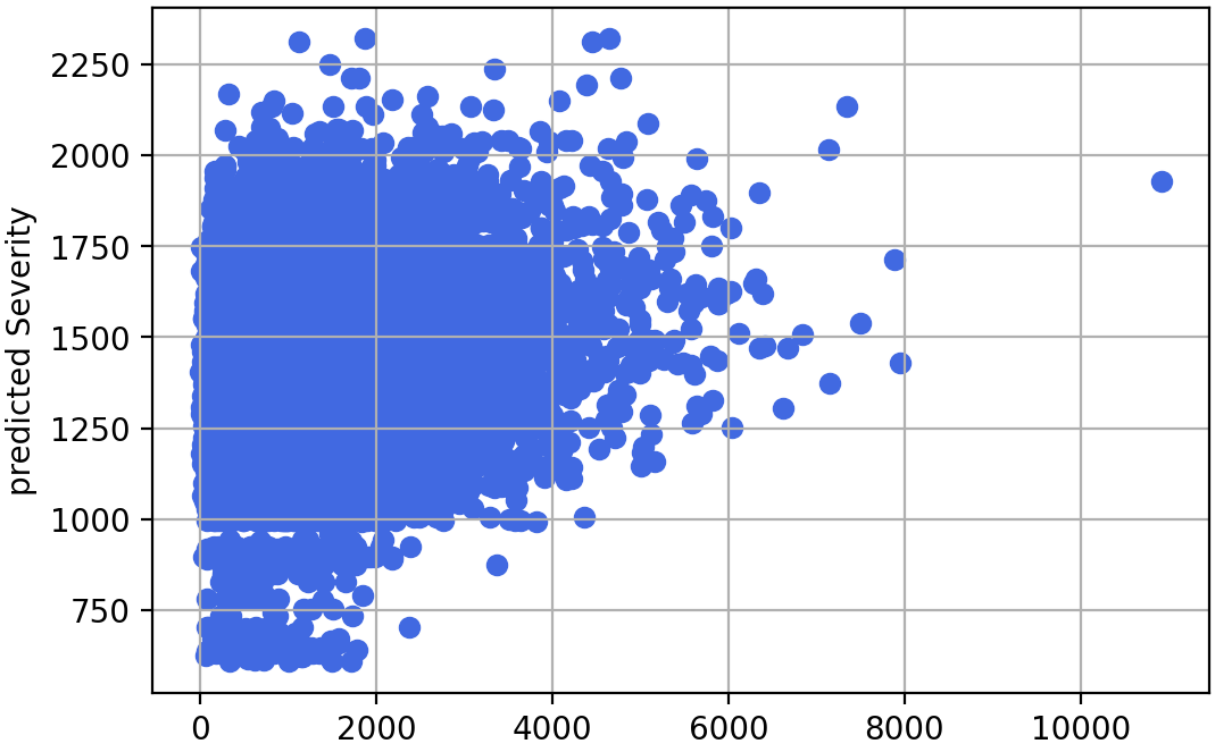
Root Mean Squared Error= 941.9847484

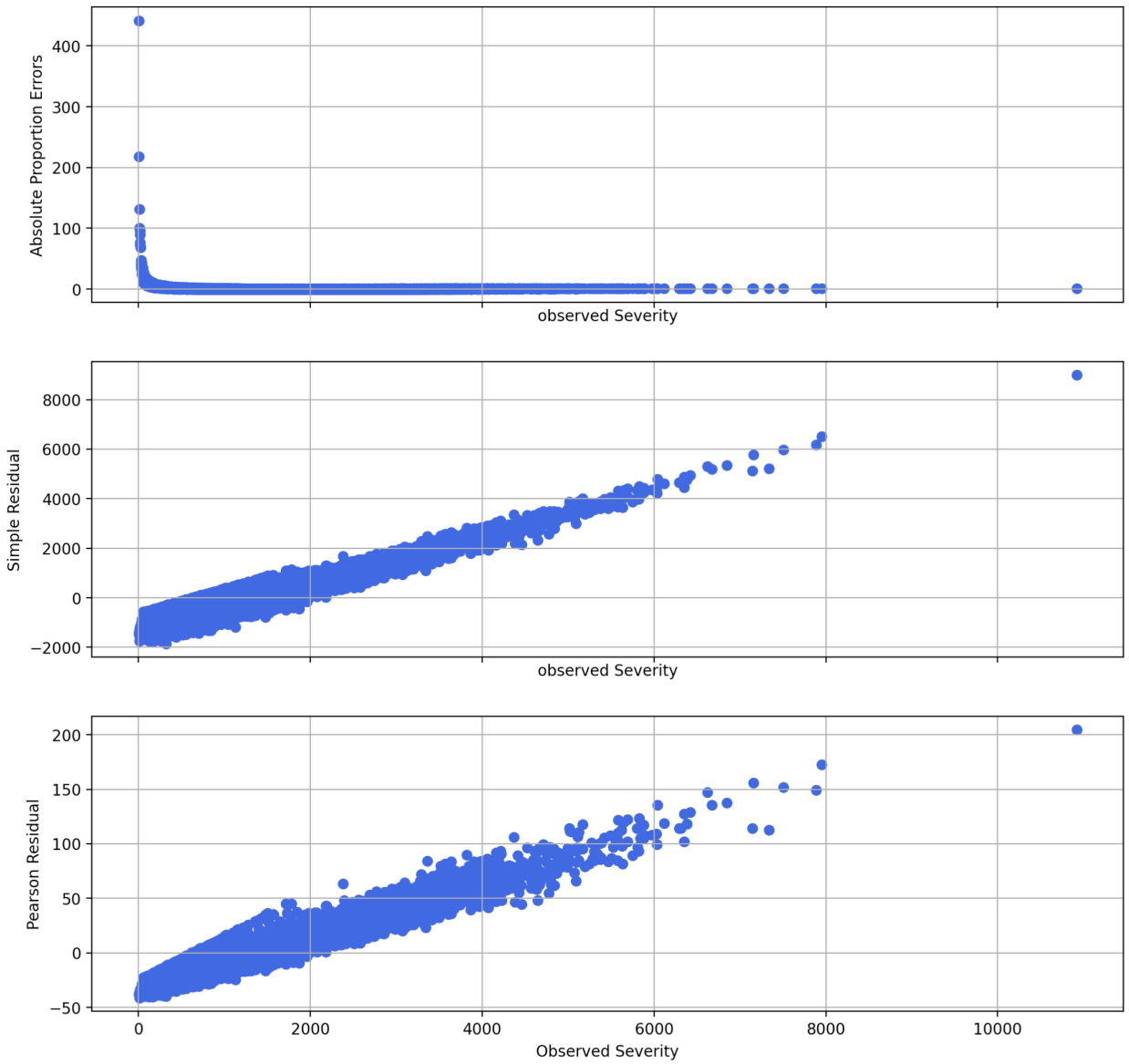
Relative Error= 0.9462485

Mean Absolute Proportion Error=1.1910876

Pearson Correlation=0.2324212

(d) (10 points) Identify any poorly predicted observations. First, plot the predicted versus the observed Severity. Second, together in a single chart frame, plot the Simple Residuals, the Pearson Residuals, and the Absolute Proportion Errors versus the observed Severity. Label the axes of these two charts accordingly. To receive full credits, generate your charts with proper dimensions (e.g., length and width) and resolution (e.g., dpi).


predicted Severity
2250
2000
1750
1500
1250
1000
750
0
2000
4000
6000
8000
10000



There is one poorly predicted observation