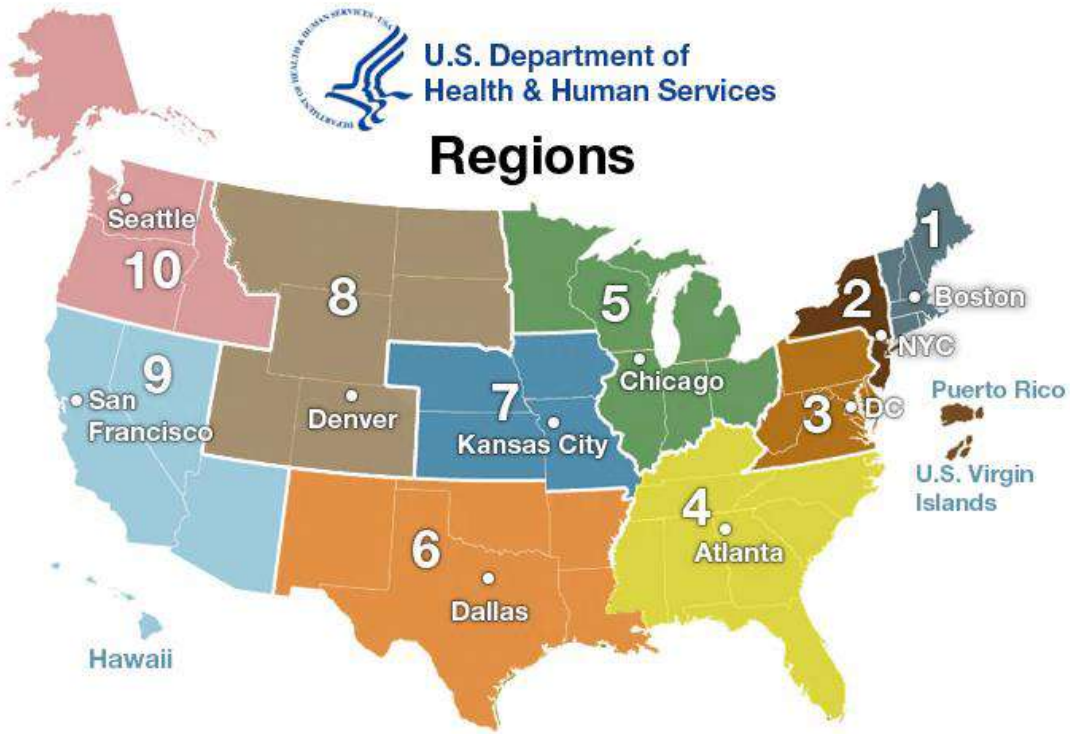
Relationship of deaths from Pneumonia and Influenza

Zhongyi Troy Zhang; Yuying Gao; Qier An
Spring 2019
MScBMI 32000

Background Information

- Influenza
- Pneumonia
- Data Resource: Deaths from Pneumonia and Influenza (P&I) and all deaths
  - Publisher: Centers for Disease Control and Prevention
  - URL:

    https://healthdata.gov/dataset/deaths-pneumonia-and-influenza-pi-and-all-deaths-state-and-region-national-center-health
- Data Cleaning



Data Manipulation

- Original headers
  - State
  - Age
  - Season
  - MMWR Year/Week
  - Deaths from Influenza
  - Deaths from pneumonia
  - All deaths
  - Percentage of deaths due to pneumonia or influenza
- Cleaned headers
  - Region
  - Age group
  - Season
  - Year
  - MMWR
  - Deaths from Influenza
  - Death from pneumonia
  - Ration between deaths

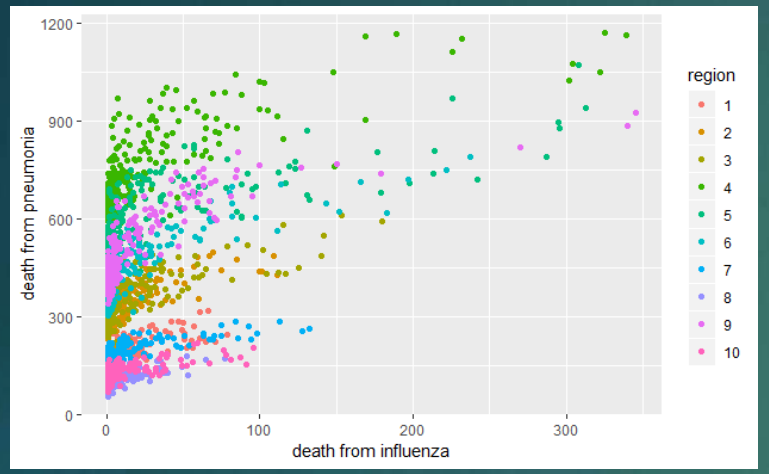
Summary Statistics (Region Chart)

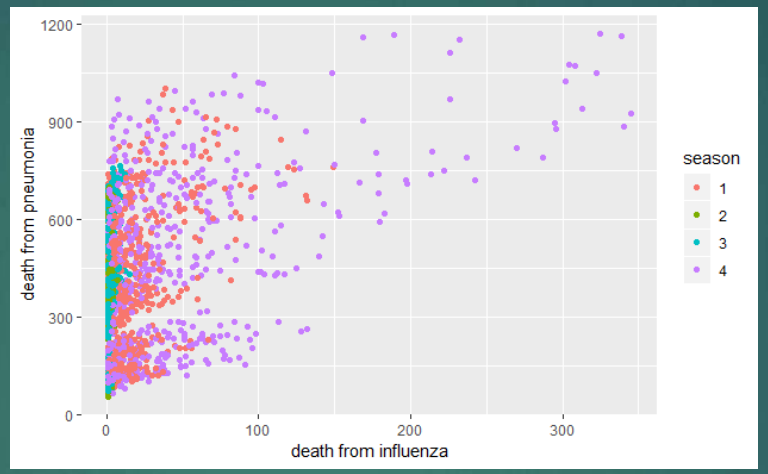
	Death from Influenza (n=38962)	Death from Pneumonia (n=908108)
Region		
1, n(%)	2424 (6.22)	46544 (5.13)
2, n(%)	2644 (6.79)	82375 (9.07)
3, n(%)	3749 (9.62)	87605 (9.65)
4, n(%)	7087 (18.19)	194014 (21.36)
5, n(%)	7193 (18.46)	152083 (16.75)
6, n(%)	4254 (10.91)	111864 (12.32)
7, n(%)	2789 (7.16)	44589 (4.91)
8, n(%)	1720 (4.41)	26647 (2.93)
9, n(%)	4794 (12.30)	130769 (14.40)
10, n(%)	2308 (5.92)	31618 (3.48)
Season		
Spring (group 1), n(%)	10704 (27.47)	239862 (26.41)
Summer (group 2), n(%)	912 (2.34)	190025 (20.93)
Fall (group 3), n(%)	1227 (3.15)	1227 (21.85)
Winter (group 4), n(%)	26119 (67.04)	26119 (30.81)
Week, beta (95% CI)	-0.85 (-0.93, -0.76)	-2.47 (-3.01, -1.92)
Year		
2014, n (%)	5511 (14.14)	178334 (19.64)
2015, n (%)	6331 (16.25)	187904 (20.69)
2016, n (%)	3831 (9.83)	178447 (19.65)
2017, n (%)	8230 (21.12)	180375 19.86)
2018, n (%)	14563 (37.38)	174586 (19.23)
2019, n (%)	496 (1.27)	8462 (0.93)

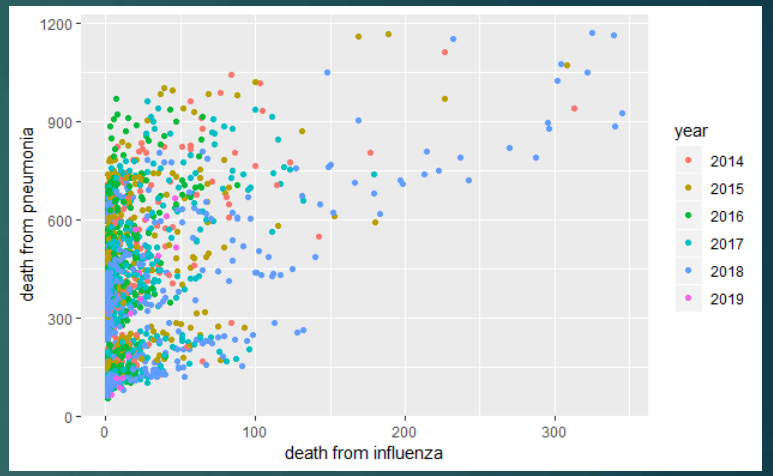
Summary Statistics (Age Chart)

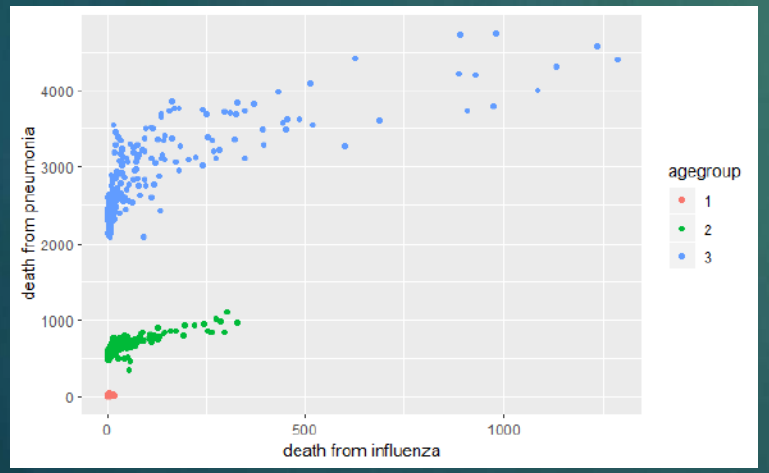
	Death from Influenza (n = 38421)	Death from Pneumonia (n = 907828)
Age		
0-17(group 1), n(%)	800 (2.08)	4001 (0.44)
18-64 (group 2), n(%)	8781 (22.85)	163269 (17.98)
over 65 (group 3), n(%)	28840 (75.06)	740558 (81.57)
Season		
Spring (group 1), n(%)	10689 (27.82)	239856 (26.42)
Summer (group 2), n(%)	618 (1.61)	189772 (20.90)
Fall (group 3), n(%)	1008 (2.62)	198413 (21.86)
Winter (group 4), n(%)	26106 (67.95)	279787 (30.82)
Week, beta (95 % CI)	-2.83 (-3.46, -2.22)	-8.23 (-14.00, -2.45)
Year		
2014, n (%)	5379 (13.81)	178323 (19.64)
2015, n (%)	6211 (15.94)	187899 (20.69)
2016, n (%)	3719 (9.55)	178442 (19.65)
2017, n (%)	8148 (20.91)	180119 (19.83)
2018, n (%)	14467 (37.13)	174583 (19.22)
2019, n (%)	497 (1.28)	8462 (0.93)

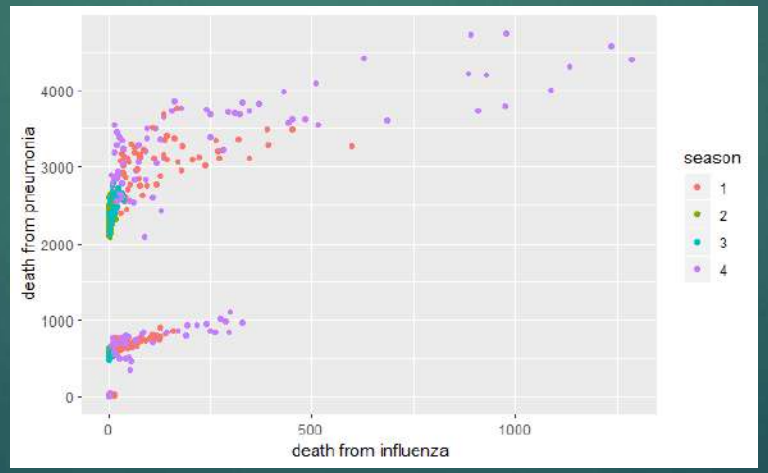
Scatter Plot

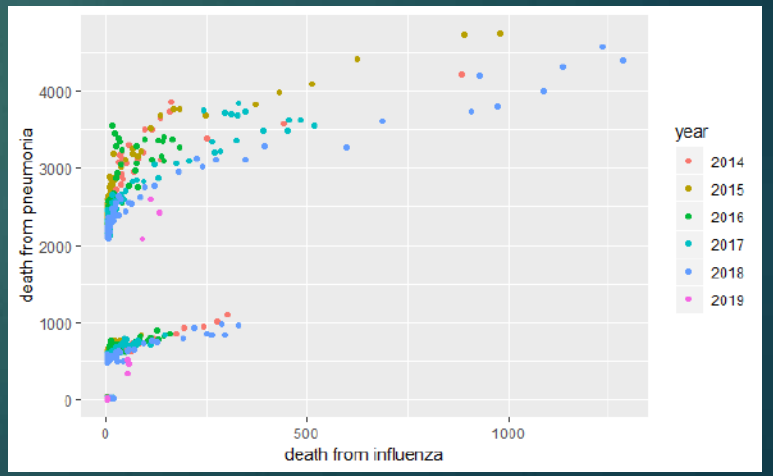












Linear Regression

Pneumonia ~ Influenza + region + season + year + week

	Beta	P-value		Beta	P-value
Intercept	211.33	< 2e-16	Season 2	-44.13	< 2e-16
Influenza	1.28	< 2e-16	Season 3	-17.89	3.63e-10 < 0.01
Region 2	134.70	< 2e-16	Season 4	33.26	< 2e-16
Region 3	150.22	< 2e-16	Year 2015	7.56	0.00247 < 0.01
Region 4	539.97	< 2e-16	Year 2016	-4.44	0.07565 > 0.01
Region 5	379.41	< 2e-16	Year 2017	-12.11	1.29e-06 < 0.01
Region 6	240.33	< 2e-16	Year 2018	-38.41	< 2e-16
Region 7	-9.25365	0.008 < 0.01	Year 2019	-143.37	< 2e-16
Region 8	-72.49	<2e-16	Week	-1.04	< 2e-16
Region 9	309.83	< 2e-16			
Region 10	-56.40	< 2e-16			

Linear Regression

Pneumonia ~ Influenza + age + season + year + week

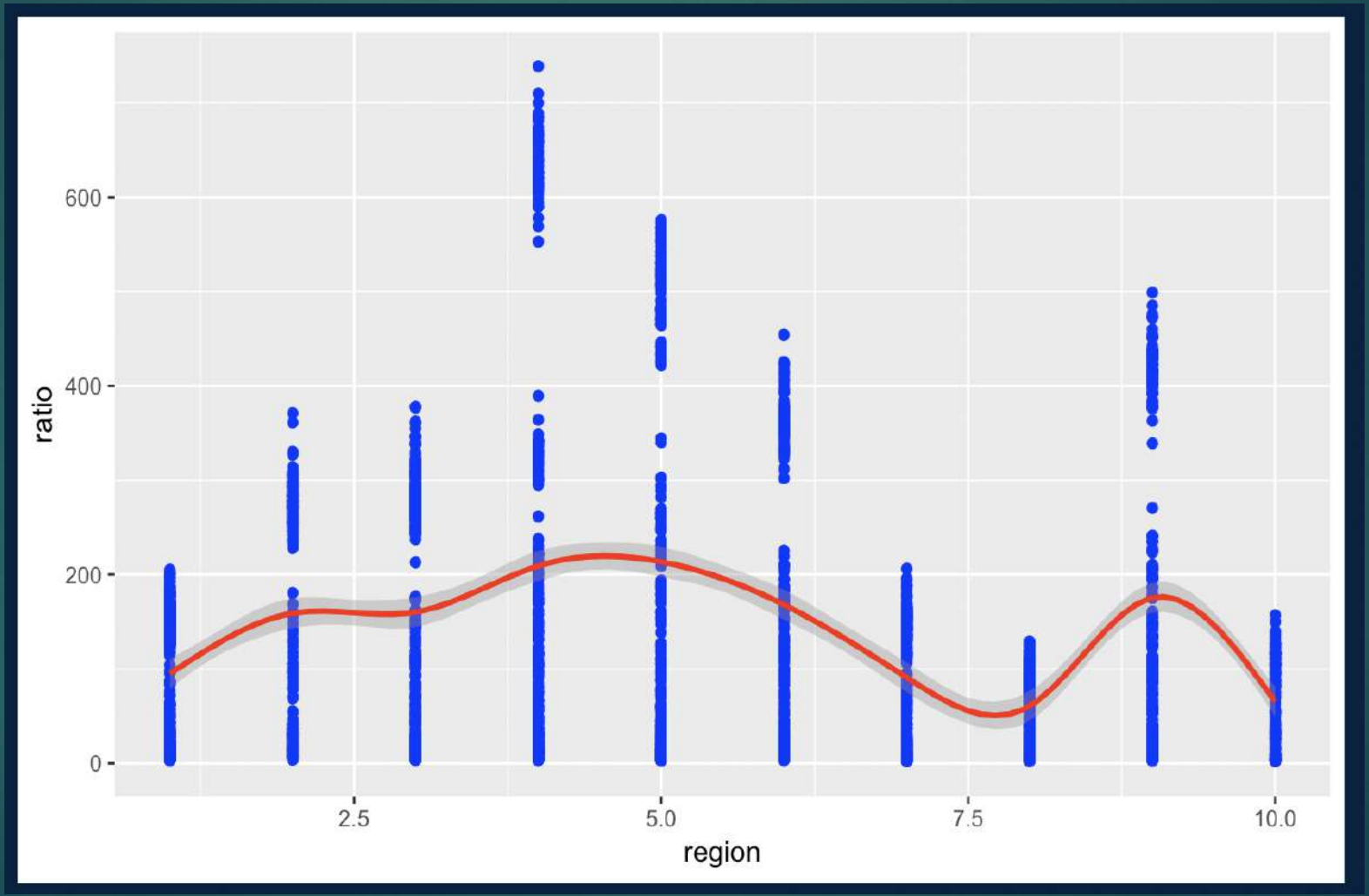
	Beta	P-value		Beta	P-value
Intercept	144.94	1.94e-11 < 0.01	Year 2015	25.60	0.188 > 0.01
Influenza	1.69	< 2e-16	Year 2016	-7.95	0.683 > 0.01
Age group 2	556.27	< 2e-16	Year 2017	-45.32	0.0199 > 0.01
Age group 3	2629.91	< 2e-16	Year 2018	-149.45	< 2e-16
Season 2	-139.46	< 2e-16	Year 2019	-424.96	<2e-16
Season 3	-68.69	0.0019 < 0.01	Week	-2.29	<2e-16
Season 4	69.18	< 2e-16			

Relationship

$$| = \frac{Death\ from\ Pneumonia}{Death\ from\ Influenza}$$

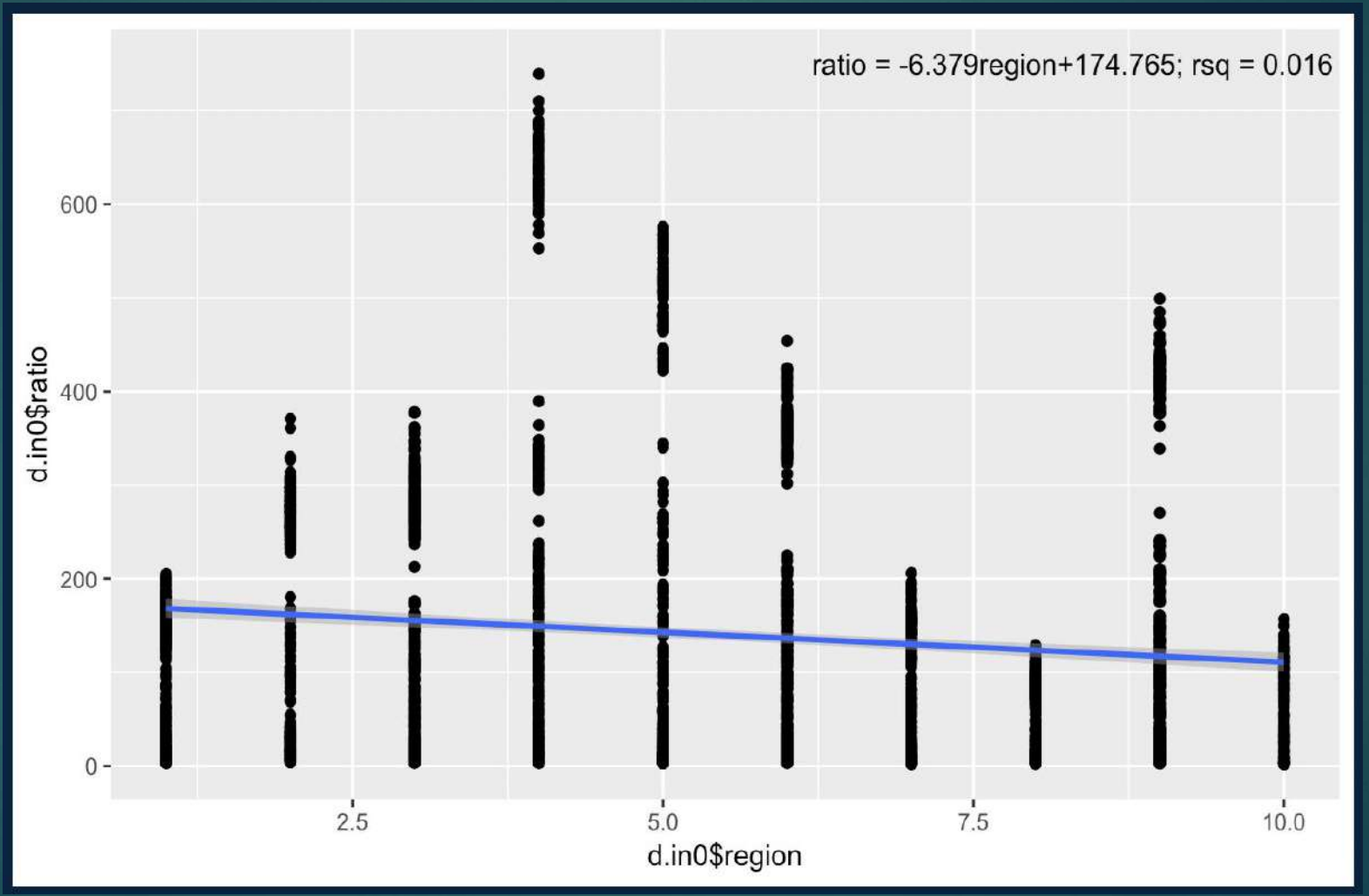
Single Linear Regression

I ~ region



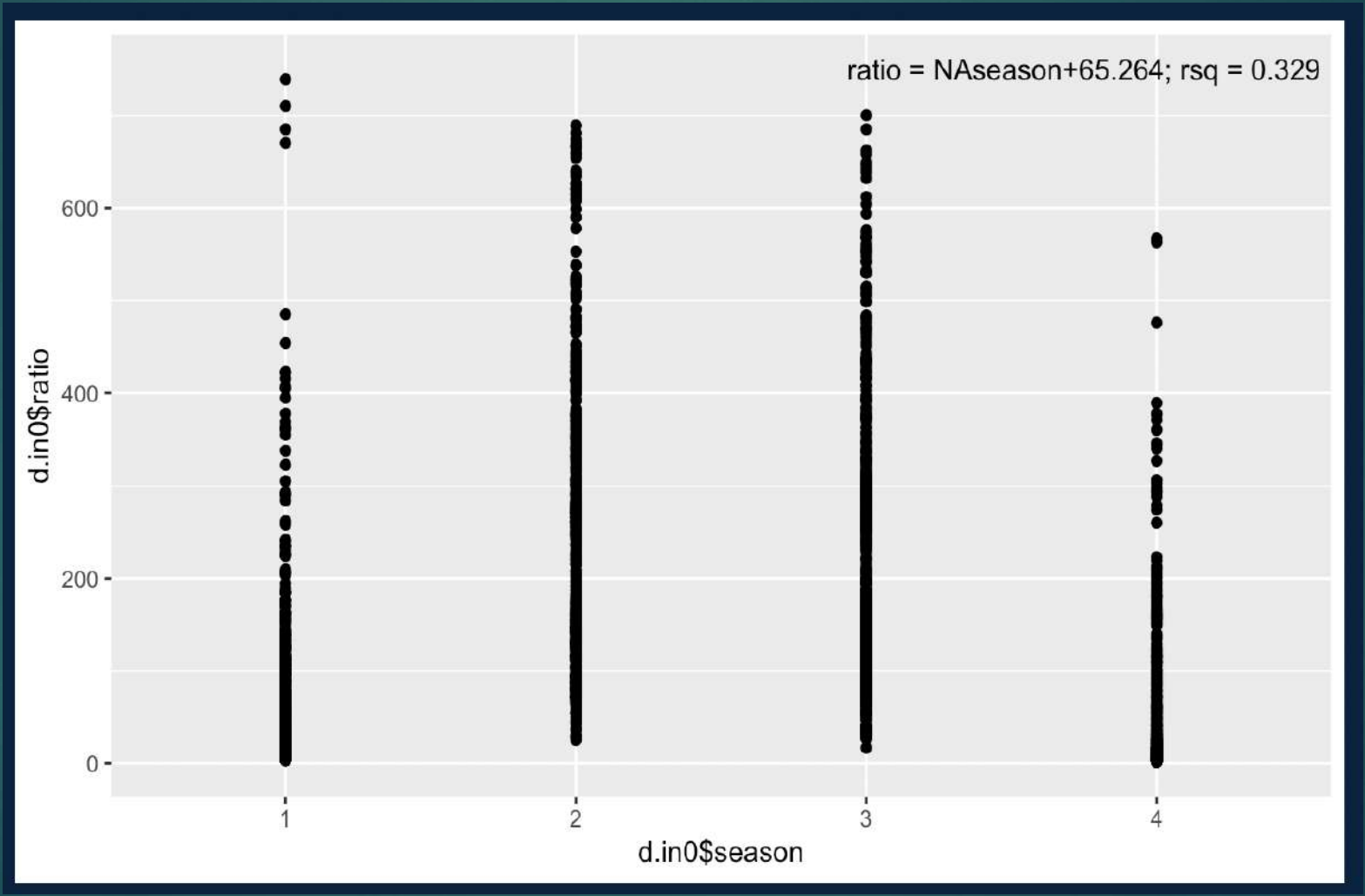
Single Linear Regression

I ~ region



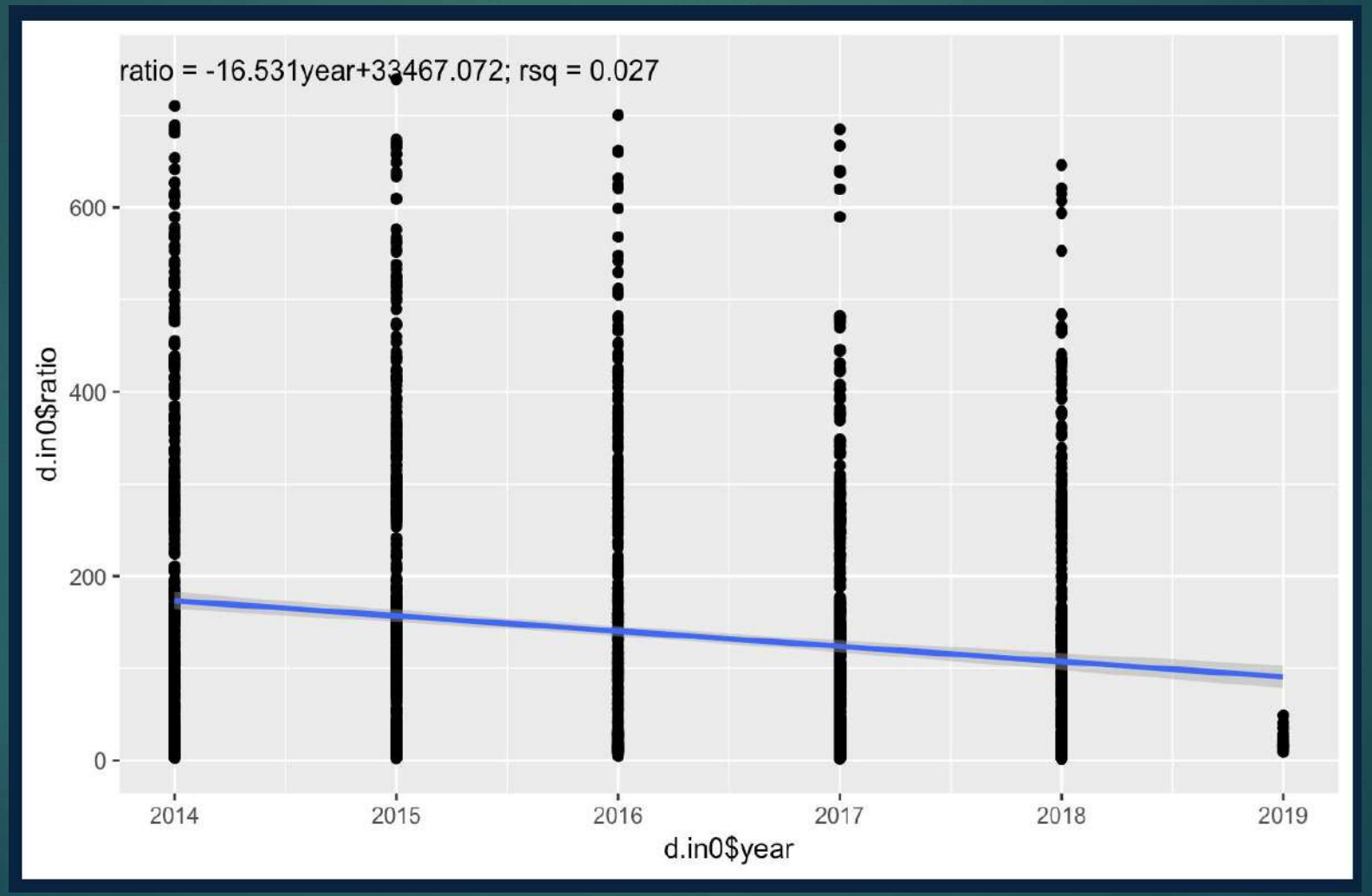
Single Linear Regression

I ~ Season



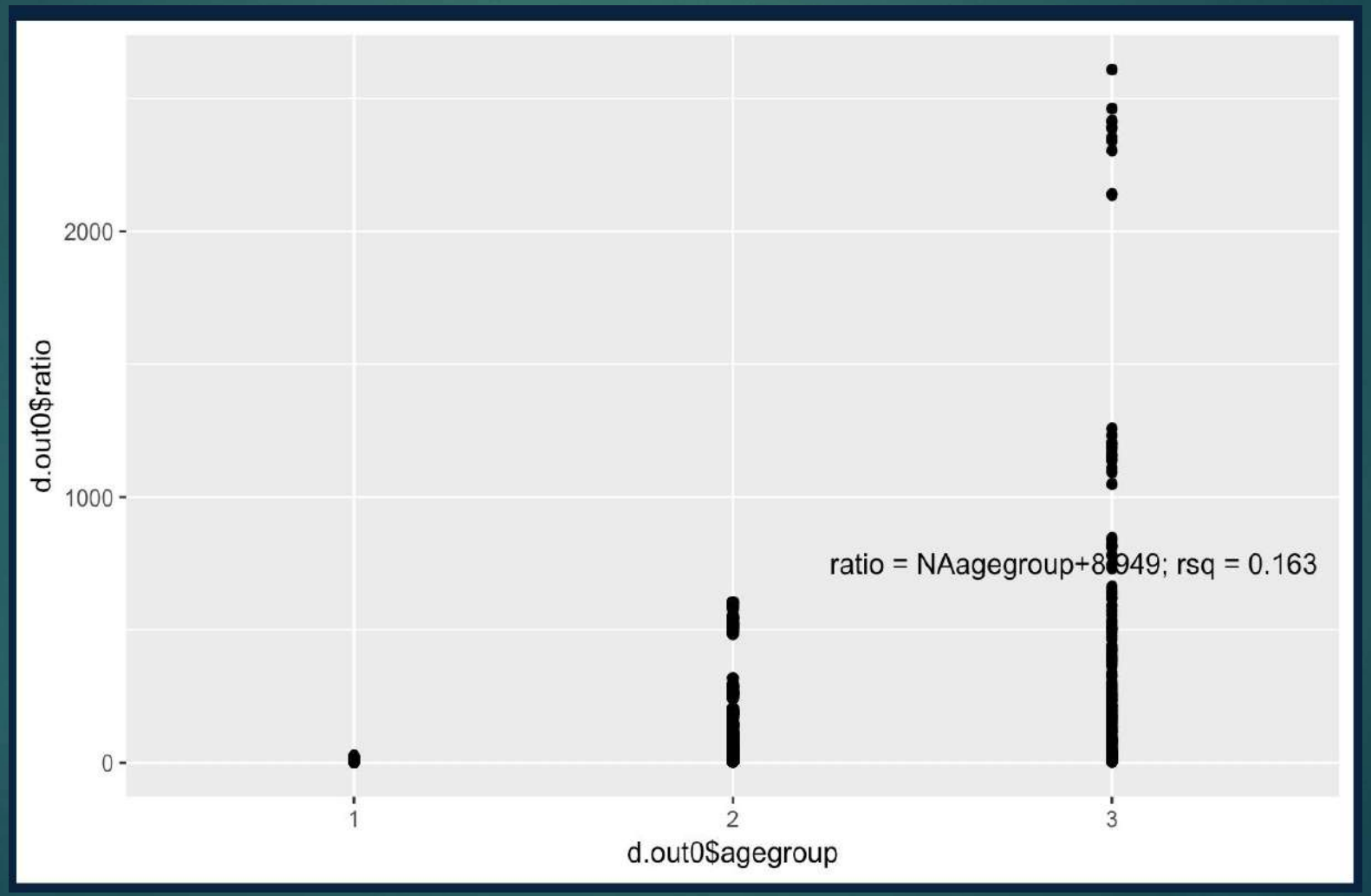
Single Linear Regression

I ~ year



Single Linear Regression

I ~ age



Linear Regression

I ~ region + season + year + week

	Beta	P-value		Beta	P-value
Intercept	19.82	0.021 > 0.01	Season 2	115.02	< 2e-16
Region 2	64.86	2.6e-13 < 0.01	Season 3	111.55	< 2e-16
Region 3	64.38	4.07e-13 < 0.01	Season 4	-21.85	0.0001 < 0.01
Region 4	115.00	< 2e-16	Year 2015	13.83	0.028 > 0.01
Region 5	118.75	< 2e-16	Year 2016	-12.39	0.049 > 0.01
Region 6	73.38	< 2e-16	Year 2017	-48.13	3.08e-14 < 0.01
Region 7	-3.36	0.703 > 0.01	Year 2018	-41.92	3.56e-11 < 0.01
Region 8	-36.54	3.62e-05 < 0.01	Year 2019	-25.91	0.188 > 0.01
Region 9	83.22	< 2e-16	Week	1.21	1.6e-12 < 0.01
Region 10	-30.46	0.00057 < 0.01			

Linear Regression

I ~ region + season + year + week

- Region 7, Year 2015, 2016 and 2019 do not have a p-value small enough to reject H0.
- These three categories do not have significant influence on the relationship between Pneumonia and Influenza
- F-statistic = 151.6 on 18 and 2602 DF, p-value < 2.2e-16
- R^2 = 0.5118
  - Adjusted R^2 = 0.5085
  - Suggesting a good fit of the model

Linear Regression

I ~ age + season + year + week

	Beta	P-value		Beta	P-value
Intercept	75.81	0.03 > 0.01	Year 2015	6.56	0.84 > 0.01
Age group 2	130.30	1.79e-07 < 0.01	Year 2016	-53.78	0.096 > 0.01
Age group 3	339.69	< 2e-16	Year 2017	-96.66	0.003 < 0.01
Season 2	317.33	< 2e-16	Year 2018	-101.03	0.002 < 0.01
Season 3	188.22	3.49e-07 < 0.01	Year 2019	-56.00	0.55 > 0.01
Season 4	-10.87	0.71 > 0.01	Week	0.43	0.62 > 0.01

Linear Regression

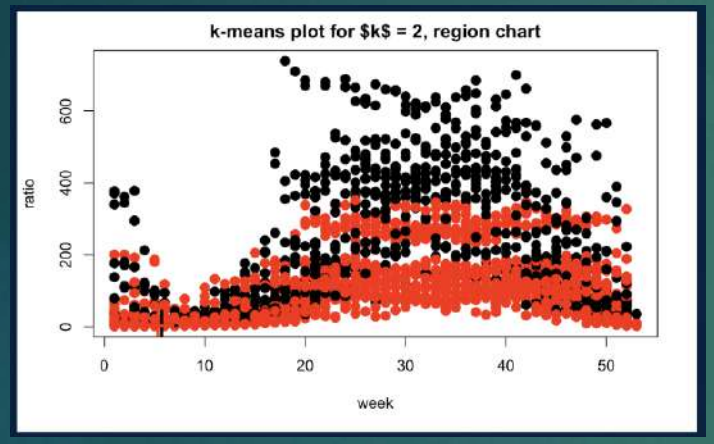
I ~ age + season + year + week

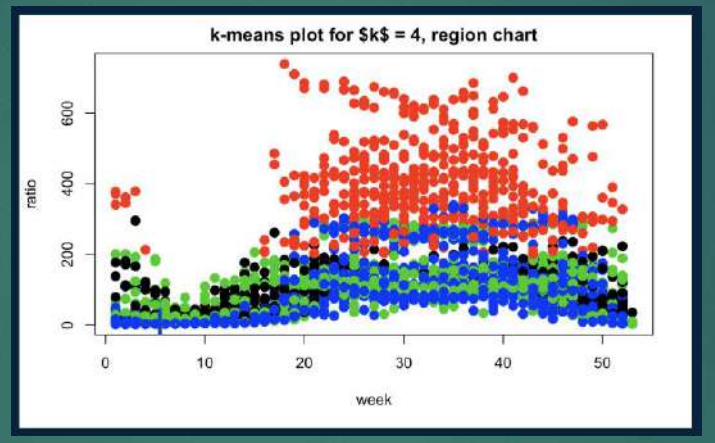
- Season 4, Year 2015, 2016 and 2019, Weeks do not have a p-value small enough to reject H0.
- These four variables do not have significant influence on the relationship between Pneumonia and Influenza
- F-statistic = 36.77 on 11 and 774 DF, p-value < 2.2e-16
- R^2 = 0.3432
  - Adjusted R^2 = 0.3339
  - Suggesting an average fit of the model

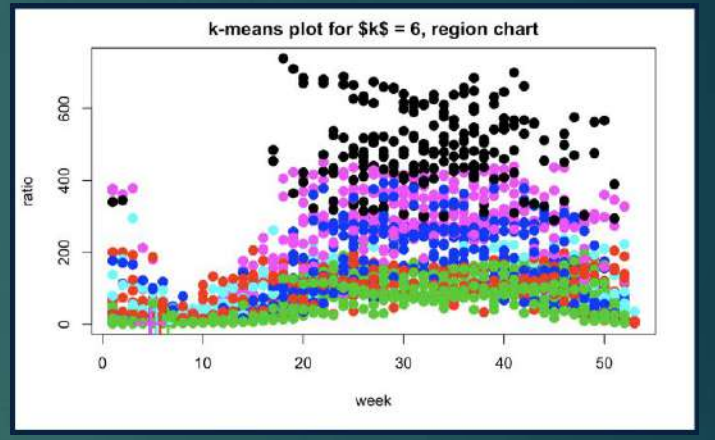
Cross Validation

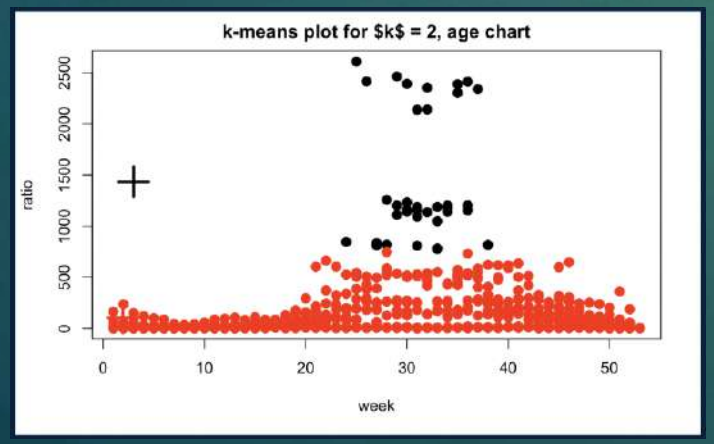
Region Chart		
Kmeans	Betweenss	Tot.withness
2	89855522	146746794
3	127413634	109188682
4	151900236	84702080
5	165076541	71525775
6	174838127	61764189
Age Chart		
Kmeans	Betweenss	Tot.withness
2	62592817	32049783
3	79085797	15556803
4	87775494	6867106
5	92157245	2485355
6	92894394	1748206

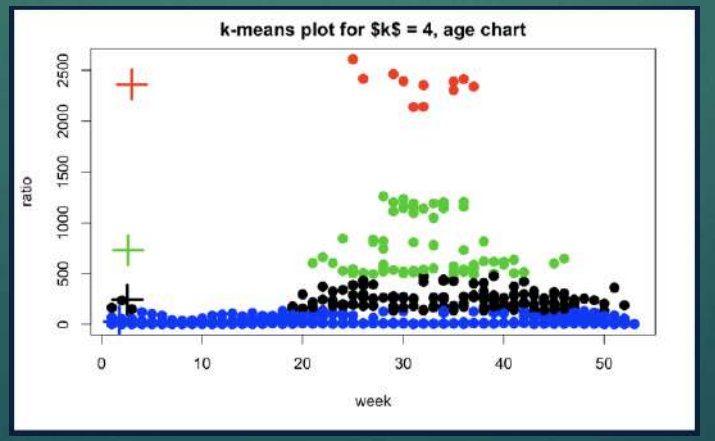
Cross Validation with different k-means

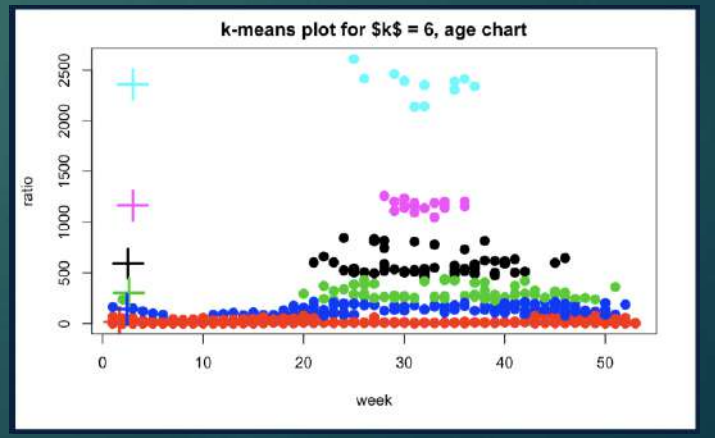












Conclusion

- Linear relationship exists between Pneumonia and Influenza
- Significant level summary
  - Region, age group, season show significant influence towards the relationship
  - Week has a moderate significant value
  - Year in general is not a significant variable for this model
- With r-square > 0.3 for model of both chart, the linear regression model can be considered a good fit
- K = 6 has the largest betweenss and smallest tot.withinss
  - A better choice

Q & A

THANK YOU