Lab 5.2 Computing the Data

Part II: Computing the Data

```
library(tidyverse)
library(stat20data)
library(Lahman)
library(broom)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ ggplot2   3.5.0     ✔ tibble    3.2.1
✔ lubridate 1.9.3     ✔ tidyr     1.3.1
✔ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
errors
```

Question 1

```
Teams_2000_present <-Teams |>
  filter(yearID >= 2000)
Teams_2000_present
```

yearID	lgID	teamID	franchID	divID	Rank	G	Ghome	W	L	⋯	DP	FP	name	park
<int>	<fct>	<fct>	<fct>	<chr>	<int>	<int>	<int>	<int>	<int>	⋯	<int>	<dbl>	<chr>	<chr>
2000	AL	ANA	ANA	W	3	162	81	82	80	⋯	182	0.978	Anaheim Angels	Edisor Intern Field
2000	NL	ARI	ARI	W	3	162	81	85	77	⋯	138	0.982	Arizona Diamondbacks	Bank (Ballpa
2000	NL	ATL	ATL	E	1	162	81	95	67	⋯	138	0.979	Atlanta Braves	Turner
2000	AL	BAL	BAL	E	4	162	81	74	88	⋯	151	0.981	Baltimore Orioles	Oriole at Can Yards
2000	AL	BOS	BOS	E	2	162	81	85	77	⋯	120	0.982	Boston Red Sox	Fenwa II
2000	AL	CHA	CHW	C	1	162	81	95	67	⋯	190	0.978	Chicago White Sox	Comis Park II

yearID	lgID	teamID	franchID	divID	Rank	G	Ghome	W	L	…	DP	FP	name	park
<int>	<fct>	<fct>	<fct>	<chr>	<int>	<int>	<int>	<int>	<int>	…	<int>	<dbl>	<chr>	<chr>
2000	NL	CHN	CHC	C	6	162	81	65	97	…	139	0.983	Chicago Cubs	Wrigle Field
2000	NL	CIN	CIN	C	2	163	82	85	77	…	156	0.982	Cincinnati Reds	Cinerg Field
2000	AL	CLE	CLE	C	2	162	81	90	72	…	147	0.988	Cleveland Indians	Jacobs
2000	NL	COL	COL	W	4	162	81	82	80	…	176	0.985	Colorado Rockies	Coors
2000	AL	DET	DET	C	3	162	81	79	83	…	171	0.983	Detroit Tigers	Come Park
2000	NL	FLO	FLA	E	3	161	81	79	82	…	144	0.980	Florida Marlins	Pro Pl Stadiu
2000	NL	HOU	HOU	C	4	162	81	72	90	…	149	0.978	Houston Astros	Enron
2000	AL	KCA	KCR	C	4	162	81	77	85	…	185	0.983	Kansas City Royals	Kauffr Stadiu
2000	NL	LAN	LAD	W	2	162	81	86	76	…	151	0.978	Los Angeles Dodgers	Dodge Stadiu
2000	NL	MIL	MIL	C	3	163	81	73	89	…	187	0.981	Milwaukee Brewers	Count Stadiu
2000	AL	MIN	MIN	C	5	162	81	69	93	…	155	0.983	Minnesota Twins	Huber Hump Metro
2000	NL	MON	WSN	E	4	162	81	67	95	…	151	0.978	Montreal Expos	Stade Olymp
2000	AL	NYA	NYY	E	1	161	80	87	74	…	132	0.981	New York Yankees	Yanke Stadiu
2000	NL	NYN	NYM	E	2	162	81	94	68	…	121	0.980	New York Mets	Shea Stadiu
2000	AL	OAK	OAK	W	1	161	81	91	70	…	164	0.978	Oakland Athletics	Oaklar Colise
2000	NL	PHI	PHI	E	5	162	81	65	97	…	136	0.983	Philadelphia Phillies	Vetera Stadiu
2000	NL	PIT	PIT	C	5	162	81	69	93	…	169	0.979	Pittsburgh Pirates	Three Stadiu
2000	NL	SDN	SDP	W	5	162	81	76	86	…	155	0.977	San Diego Padres	Qualc Stadiu
2000	AL	SEA	SEA	W	2	162	81	91	71	…	176	0.984	Seattle Mariners	Safeco

yearID	lgID	teamID	franchID	divID	Rank	G	Ghome	W	L	…	DP	FP	name	park
<int>	<fct>	<fct>	<fct>	<chr>	<int>	<int>	<int>	<int>	<int>	…	<int>	<dbl>	<chr>	<chr
2000	NL	SFN	SFG	W	1	162	81	97	65	…	173	0.985	San Francisco Giants	PacBe
2000	NL	SLN	STL	C	1	162	81	95	67	…	148	0.981	St. Louis Cardinals	Busch Stadiu
2000	AL	TBA	TBD	E	5	161	80	69	92	…	169	0.981	Tampa Bay Devil Rays	Tropic Field
2000	AL	TEX	TEX	W	4	162	81	71	91	…	162	0.978	Texas Rangers	The Ba at Arli
2000	AL	TOR	TOR	E	3	162	81	83	79	…	176	0.984	Toronto Blue Jays	Skydo
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋮
2022	NL	ARI	ARI	W	4	162	81	74	88	…	134	0.985	Arizona Diamondbacks	Chase
2022	NL	ATL	ATL	E	1	162	81	101	61	…	110	0.987	Atlanta Braves	SunTru Park
2022	AL	BAL	BAL	E	4	162	81	83	79	…	151	0.985	Baltimore Orioles	Oriole at Can Yards
2022	AL	BOS	BOS	E	5	162	81	78	84	…	134	0.985	Boston Red Sox	Fenwa II
2022	AL	CHA	CHW	C	2	162	81	81	81	…	122	0.982	Chicago White Sox	Guara Rate F
2022	NL	CHN	CHC	C	3	162	81	74	88	…	139	0.984	Chicago Cubs	Wrigle Field
2022	NL	CIN	CIN	C	4	162	81	62	100	…	115	0.986	Cincinnati Reds	Great Ameri Ball Pa
2022	AL	CLE	CLE	C	1	162	81	92	70	…	127	0.984	Cleveland Guardians	Progre Field
2022	NL	COL	COL	W	5	162	81	68	94	…	154	0.983	Colorado Rockies	Coors
2022	AL	DET	DET	C	4	162	82	66	96	…	137	0.984	Detroit Tigers	Come Park
2022	AL	HOU	HOU	W	1	162	81	106	56	…	122	0.987	Houston Astros	Minut Maid I
2022	AL	KCA	KCR	C	5	162	81	65	97	…	153	0.986	Kansas City Royals	Kauffr Stadiu
2022	AL	LAA	ANA	W	3	162	81	73	89	…	134	0.985	Los Angeles Angels of	Angel Stadiu

yearID	lgID	teamID	franchID	divID	Rank	G	Ghome	W	L	⋯	DP	FP	name	park
<int>	<fct>	<fct>	<fct>	<chr>	<int>	<int>	<int>	<int>	<int>	⋯	<int>	<dbl>	<chr>	<chr>
													Anaheim	Anahe
2022	NL	LAN	LAD	W	1	162	81	111	51	⋯	120	0.986	Los Angeles Dodgers	Dodge Stadiu
2022	NL	MIA	FLA	E	4	162	81	69	93	⋯	143	0.988	Miami Marlins	Marlin
2022	NL	MIL	MIL	C	2	162	81	86	76	⋯	122	0.984	Milwaukee Brewers	Miller
2022	AL	MIN	MIN	C	3	162	81	78	84	⋯	121	0.985	Minnesota Twins	Target
2022	AL	NYA	NYY	E	1	162	81	99	63	⋯	102	0.987	New York Yankees	Yanke Stadiu
2022	NL	NYN	NYM	E	2	162	81	101	61	⋯	128	0.988	New York Mets	Citi Fie
2022	AL	OAK	OAK	W	5	162	80	60	102	⋯	139	0.984	Oakland Athletics	O.co Colise
2022	NL	PHI	PHI	E	3	162	81	87	75	⋯	129	0.988	Philadelphia Phillies	Citizer Bank F
2022	NL	PIT	PIT	C	5	162	81	62	100	⋯	152	0.979	Pittsburgh Pirates	PNC P
2022	NL	SDN	SDP	W	2	162	81	89	73	⋯	116	0.987	San Diego Padres	Petco
2022	AL	SEA	SEA	W	2	162	81	90	72	⋯	114	0.988	Seattle Mariners	T-Mob Park
2022	NL	SFN	SFG	W	3	162	81	81	81	⋯	130	0.983	San Francisco Giants	Oracle
2022	NL	SLN	STL	C	1	162	81	93	69	⋯	181	0.989	St. Louis Cardinals	Busch Stadiu
2022	AL	TBA	TBD	E	1	162	81	86	76	⋯	110	0.985	Tampa Bay Rays	Tropic Field
2022	AL	TEX	TEX	W	4	162	81	68	94	⋯	143	0.984	Texas Rangers	Globe Field
2022	AL	TOR	TOR	E	2	162	81	92	70	⋯	120	0.986	Toronto Blue Jays	Roger Centre
2022	NL	WAS	WSN	E	5	162	81	55	107	⋯	126	0.982	Washington Nationals	Natior Park

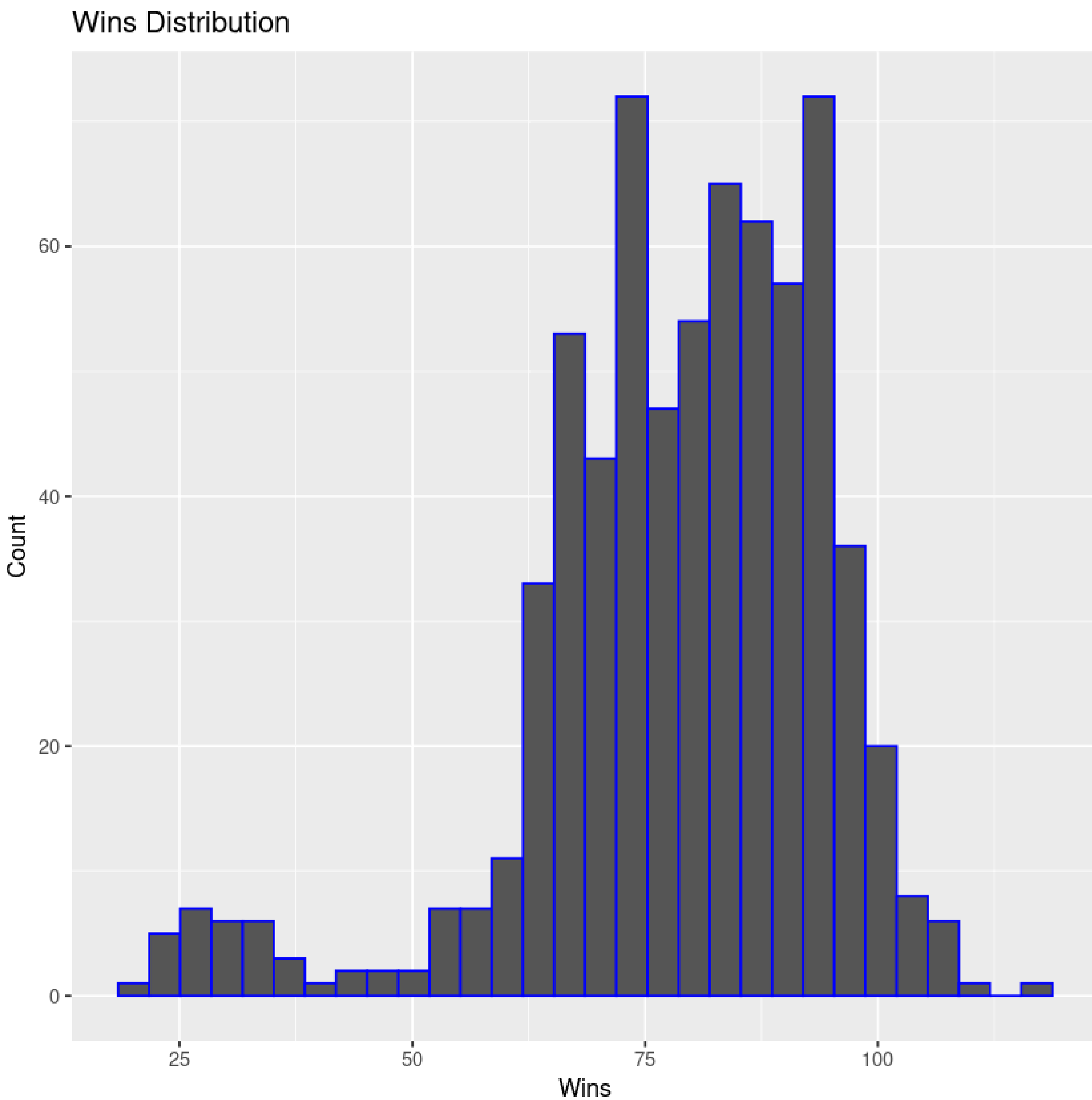
A data.frame: 690 × 48

The dimensions for the filtered data set from 2000 to the present are 690 x 48.

Question 2

```
Teams_2000_present|>
  ggplot(aes(x = W)) +
  geom_histogram(color = "blue") +
  labs(x = "Wins", y = "Count") +
  ggtitle("Wins Distribution")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The shape of the win distribution is bimodal and left-skewed. The distribution shown was similar to what I had in my speculations from part 1. Since there are 162 baseball games per year, I expected that most teams would win about 40% - 60% of their games in each season and the data accurately predicted my speculation.

Question 3

```
Teams_2000_present|>
  ggplot(aes(x = R,
             y = W)) +
  geom_point(color = "red") +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Runs", y = "Wins") +
  ggtitle("Runs and Wins Distribution")

Teams_2000_present |>
  filter(W < 60, R < 400)
```

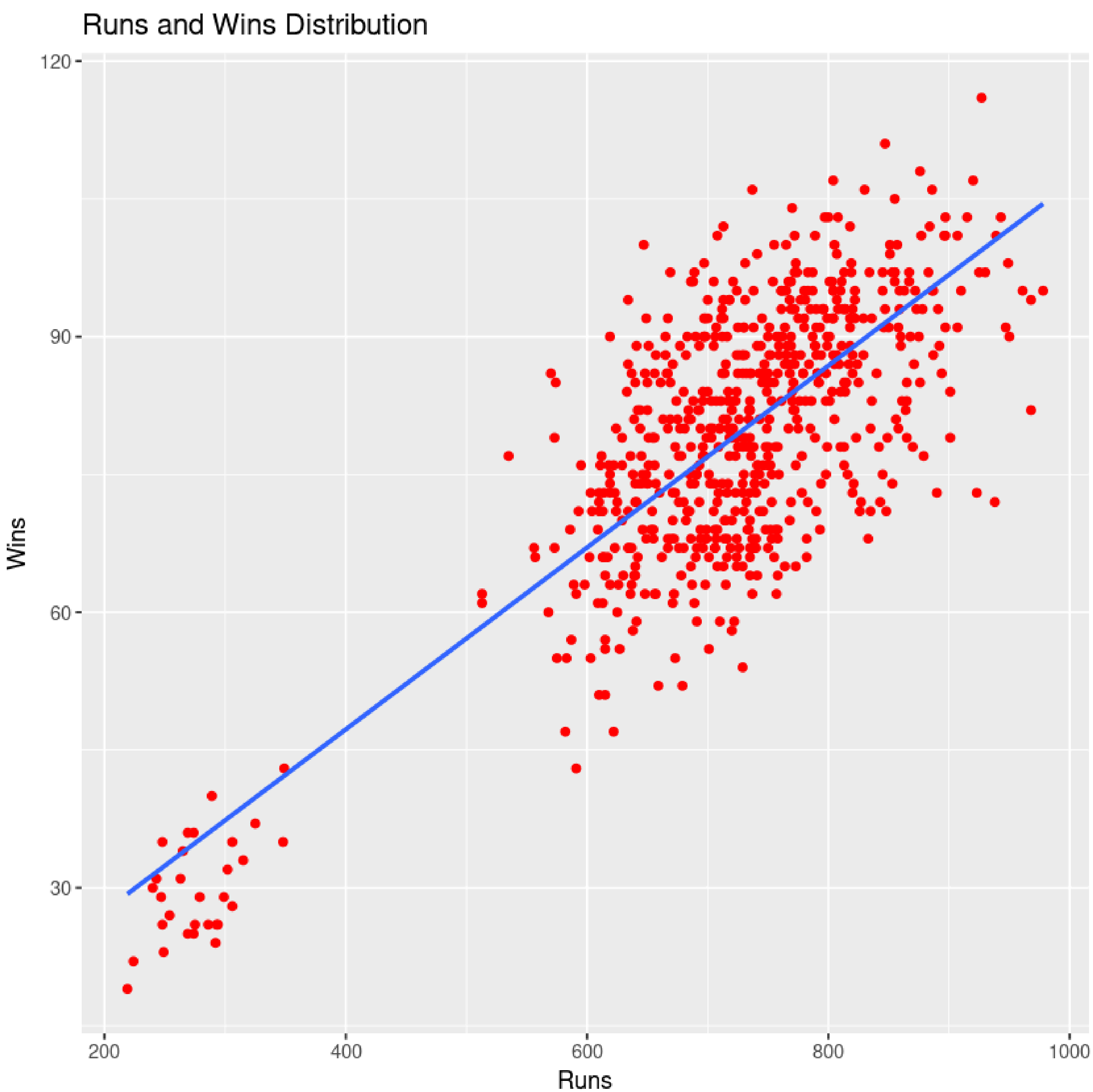
```
`geom_smooth()` using formula = 'y ~ x'
```

yearID	lgID	teamID	franchID	divID	Rank	G	Ghome	W	L	⋯	DP	FP	name	park
<int>	<fct>	<fct>	<fct>	<chr>	<int>	<int>	<int>	<int>	<int>	⋯	<int>	<dbl>	<chr>	<chr>
2020	NL	ARI	ARI	W	5	60	30	25	35	⋯	54	0.983	Arizona Diamondbacks	Chase
2020	NL	ATL	ATL	E	1	60	30	35	25	⋯	52	0.985	Atlanta Braves	SunTru Park
2020	AL	BAL	BAL	E	4	60	33	25	35	⋯	42	0.980	Baltimore Orioles	Oriole at Can Yards
2020	AL	BOS	BOS	E	5	60	31	24	36	⋯	59	0.979	Boston Red Sox	Fenwa Park II
2020	AL	CHA	CHW	C	2	60	30	35	25	⋯	48	0.982	Chicago White Sox	Guara Rate F
2020	NL	CHN	CHC	C	1	60	33	34	26	⋯	46	0.986	Chicago Cubs	Wrigle Field
2020	NL	CIN	CIN	C	2	60	29	31	29	⋯	36	0.986	Cincinnati Reds	Great Ameri Ball Pa
2020	AL	CLE	CLE	C	2	60	30	35	25	⋯	46	0.986	Cleveland Indians	Progre Field
2020	NL	COL	COL	W	4	60	30	26	34	⋯	78	0.981	Colorado Rockies	Coors
2020	AL	DET	DET	C	5	58	27	23	35	⋯	46	0.987	Detroit Tigers	Come Park

yearID	lgID	teamID	franchID	divID	Rank	G	Ghome	W	L	⋯	DP	FP	name	park
<int>	<fct>	<fct>	<fct>	<chr>	<int>	<int>	<int>	<int>	<int>	⋯	<int>	<dbl>	<chr>	<chr>
2020	AL	HOU	HOU	W	2	60	28	29	31	⋯	48	0.991	Houston Astros	Minut Maid I
2020	AL	KCA	KCR	C	4	60	30	26	34	⋯	62	0.985	Kansas City Royals	Kauffr Stadiu
2020	AL	LAA	ANA	W	4	60	31	26	34	⋯	36	0.983	Los Angeles Angels of Anaheim	Angel Stadiu Anahe
2020	NL	LAN	LAD	W	1	60	30	43	17	⋯	46	0.982	Los Angeles Dodgers	Dodge Stadiu
2020	NL	MIA	FLA	E	2	60	26	31	29	⋯	60	0.981	Miami Marlins	Marlin Park
2020	NL	MIL	MIL	C	4	60	29	29	31	⋯	45	0.984	Milwaukee Brewers	Miller
2020	AL	MIN	MIN	C	1	60	31	36	24	⋯	39	0.990	Minnesota Twins	Target
2020	AL	NYA	NYY	E	2	60	31	33	27	⋯	37	0.976	New York Yankees	Yanke Stadiu
2020	NL	NYN	NYM	E	4	60	29	26	34	⋯	39	0.985	New York Mets	Citi Fie
2020	AL	OAK	OAK	W	1	60	32	36	24	⋯	33	0.987	Oakland Athletics	O.co Colise
2020	NL	PHI	PHI	E	3	60	32	28	32	⋯	57	0.983	Philadelphia Phillies	Citizer Bank F
2020	NL	PIT	PIT	C	5	60	32	19	41	⋯	53	0.978	Pittsburgh Pirates	PNC P
2020	NL	SDN	SDP	W	2	60	32	37	23	⋯	46	0.985	San Diego Padres	Petco
2020	AL	SEA	SEA	W	3	60	24	27	33	⋯	48	0.989	Seattle Mariners	T-Mob Park
2020	NL	SFN	SFG	W	3	60	33	29	31	⋯	43	0.980	San Francisco Giants	Oracle
2020	NL	SLN	STL	C	3	58	27	30	28	⋯	46	0.983	St. Louis Cardinals	Busch Stadiu
2020	AL	TBA	TBD	E	1	60	29	40	20	⋯	52	0.985	Tampa Bay Rays	Tropic Field
2020	AL	TEX	TEX	W	5	60	30	22	38	⋯	40	0.981	Texas Rangers	Globe Field
2020	AL	TOR	TOR	E	3	60	26	32	28	⋯	47	0.982	Toronto Blue Jays	Sahler Field

yearID	lgID	teamID	franchID	divID	Rank	G	Ghome	W	L	⋯	DP	FP	name	park
<int>	<fct>	<fct>	<fct>	<chr>	<int>	<int>	<int>	<int>	<int>	⋯	<int>	<dbl>	<chr>	<chr>
2020	NL	WAS	WSN	E	4	60	33	26	34	⋯	48	0.981	Washington Nationals	Natior Park

A data.frame: 30 × 48

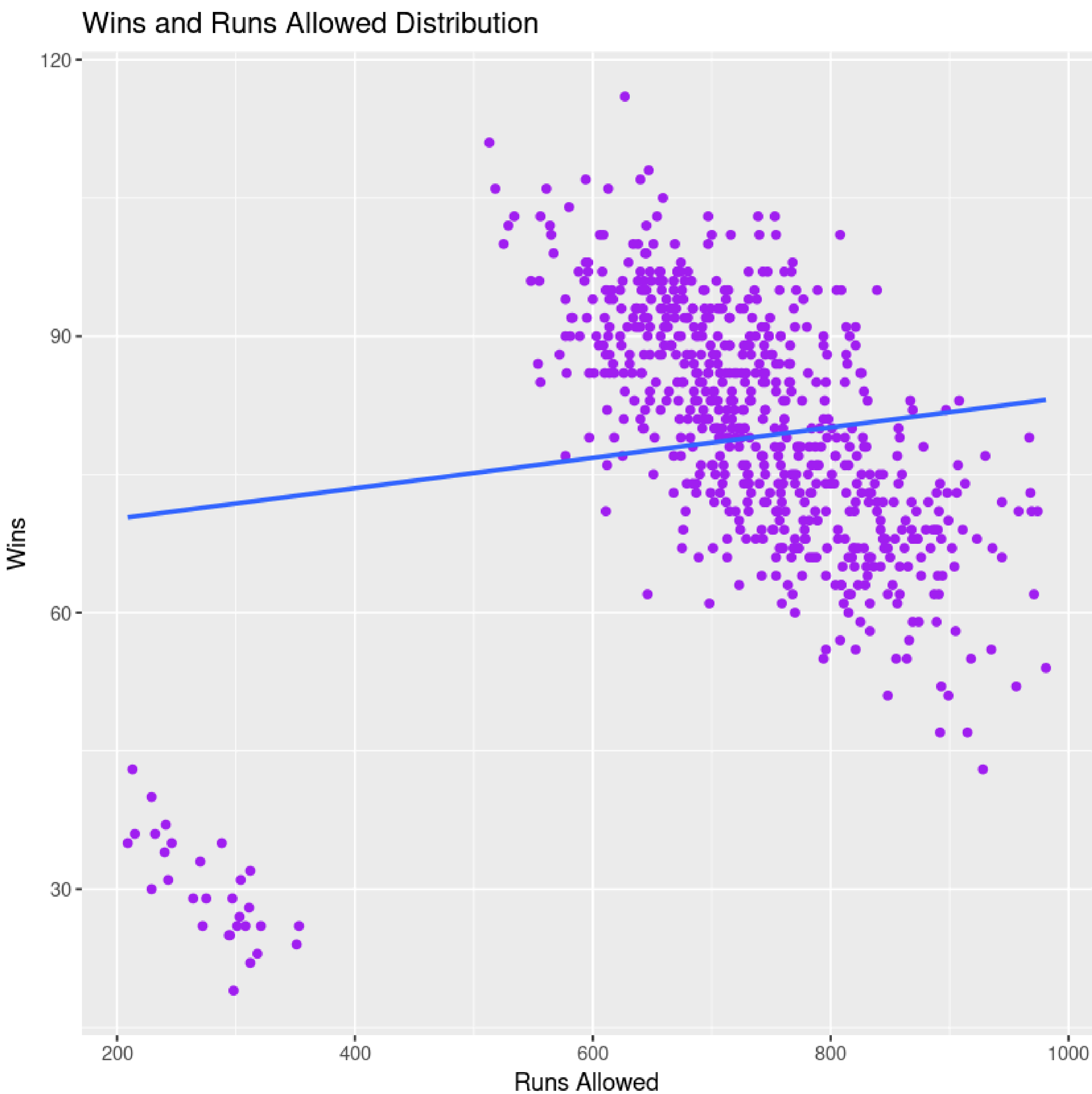


The Wins and Runs Distribution are linear and seem to have a linear and very strong positive correlation on a scatter plot. There are two clusters of data formed closely together, one larger than the other. The reason for the smaller cluster was that during the COVID-19 pandemic in 2020, many teams played only 60 or fewer games, with the rest of the games being canceled.

Question 4

```
Teams_2000_present |>
  ggplot(aes(x = RA,
             y = W)) +
  geom_point(color = "purple") +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Runs Allowed", y = "Wins") +
  ggtitle("Wins and Runs Allowed Distribution")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



The Wins and Runs Allowed Distribution seems to have a linear and very strong negative correlation on a scatter plot. The distribution also has two clusters the Wins and Runs Allowed Distribution where the small cluster represents fewer games being played in 2020. Due to the small cluster of outliers representing fewer games played in the 2020 season, the least squares line is affected and may not accurately represent the model. While the Runs and Wins Distribution is positive and the Runs Allowed and Wins Distribution is negative, their correlational strength seem similar.

Question 5

```
model_1 <- lm(formula = W ~ R, data = Teams_2000_present)
model_1

glance(model_1)|>
  select(r.squared)
```

```
Call:
lm(formula = W ~ R, data = Teams_2000_present)

Coefficients:
(Intercept)            R
    7.65895      0.09899
```

r.squared
<dbl>
0.6071875

A tibble: 1 × 1

Linear Model Equation: Wins(hat) = 7.65895 + 0.09899 x Runs.

The R squared in the context of the problem will be used to predict the number of wins using the number of runs. The R squared of approximately 61% represents that about 61% of the number of runs can explain about 61% of the variability found in the number of wins. This means the number of runs is a good, but not the best indicator for the number of wins.

Question 6

```
Teams_2000_present|>
  summarise(runs_mean = mean(R),
            wins_mean = mean(W))
```

runs_mean	wins_mean
<dbl>	<dbl>
718.2203	78.75217

A data.frame: 1 × 2

The average number of season runs is 718.2203 and the average number of season wins is 78.75217. Using our linear model equation from Question 5, we would input 718.2203 into the equation as the number of runs: Wins(hat) = 7.65895 + 0.09899 x Runs.

```
7.65895 + 0.09899 * (718.75217)
```

78.8082273083
I would predict a team that scored 718.7512 runs to win 78.8082273083, or about 79 games in a single season, which is about 49% of the 162 games played in a single season.

Question 7

Using the same equation, we can predict the number of games for a team that scored 600, 850, and 10,000 runs.

```
7.65895 + 0.09899 * (600)
```

67.05295
A team that has scored 600 runs is predicted to win 67.05295 games or about 67 games.

```
7.65895 + 0.09899 * (850)
```

91.80045
A team that has scored 850 runs is predicted to win 91.80045 games or about 92 games.

```
7.65895 + 0.09899 * (10000)
```

997.55895
A team that has scored 10000 runs is predicted to win 997.55895, or about 998 games.

The 10,000 runs prediction is inaccurate for our linear model because the typical baseball game is 162 games. A team can’t play 998 games in a single season, making the 10,000 runs prediction inaccurate. Also, due to the small clusters of outliers from games played in 2020, our predicted number of wins from 600, 850, and 10000 runs may be slightly inaccurate as opposed to if 162 games were typically played in 2020.

Question 8

```r
model_2 <- lm(formula = W ~ R + RA, data = Teams_2000_present)
model_2

glance(model_2)|>
  select(r.squared)
```

```
Call:
lm(formula = W ~ R + RA, data = Teams_2000_present)

Coefficients:
(Intercept)            R           RA
    25.0971       0.1400      -0.0653
```

r.squared
<dbl>
0.786975

A tibble: 1 × 1

Equation for Multiple Linear Regression Model: Wins(hat) = 25.0971 + 0.14 x Runs - 0.0653 x Runs Allowed.

The R squared for this multiple linear regression model is 0.789675. This model has a higher R squared value than the previous model, which makes it a better predictor for the number of season wins. Adding Runs Allowed as another predictor variable will increase R squared as a result.

Question 9

```r
Teams_2000_present <- Teams_2000_present |>
    mutate(log_runs = log(R))|>
    mutate(log_RA = log(RA))|>
    mutate(log_doubles = log(X2B))|>
    mutate(log_saves = log(SV))

model_3 <- lm(formula = W ~ log_runs + log_RA + log_doubles + log_saves, data = Teams_20
model_3

glance(model_3)|>
  select(r.squared)
```

```
Call:
lm(formula = W ~ log_runs + log_RA + log_doubles + log_saves,
    data = Teams_2000_present)

Coefficients:
```

```
(Intercept)     log_runs       log_RA  log_doubles    log_saves
    -156.84        65.03       -46.66         5.91        22.51
```

r.squared
<dbl>
0.9350519

A tibble: 1 × 1

For my model, I used the log of runs, the log of runs allowed, the log of doubles, and the log of saves as variables to predict the number of wins.

The equation of my resulting linear model is: Wins(hat) = -156.84 + 65.03 x log(Runs) - 46.66 x log(Runs Allowed) + 5.91 x log(Doubles) + 22.51 x log(Saves)

The R squared for my resulting linear model is 0.9350519. Since this model has two new predictor variables and three non-linear transformations, this makes this model the best predictor for the number of seasons compared to the other two graphs.

Question 10

Causation is when one variable Causation is when one variable directly influences the change in another variable, in other words is a cause-and-effect relationship. However, correlation does not imply causation. Even if my predictor variable has a positive coefficient in my predictive model, a sports management team and I cannot imply causation based on the coefficients from experimental studies, due to confounding variables. To make causal claims, I would need to gather data from observational studies.