

UNIVERSIDAD ICESI

Análisis Exploratorio de Datos

PROYECTO FINAL

Autores:

María Paula Fernández

Cristian Palechor

Jorge Antonio Morales

Facultad de Ingeniería

Maestría en Ciencia de Datos

Marzo de 2022

Introducción (1RA VERSION)

En el primer cuarto del año 2020 todo el mundo se vio afectado por virus sarc covid 19, paralizando la economía, eventos y demás situaciones “normales” a las cuales estábamos acostumbrados. Los eventos masivos se vieron fuertemente involucrados a pasar de un aforo 100% de capacidad a restringir por completo la capacidad de ocupación de los eventos.

El objetivo del proyecto de analítica es revisar las bases de datos suministradas de la liga inglesa de futbol y analizar si se afecto la ventaja que tiene el equipo local al jugar en “casa” a puerta cerrada.

Sobre los datos

```
inglesa_final$FTHG = ifelse(inglesa_final$FTHG>=4, "4 o más", inglesa_final$FTHG)
inglesa_final$FTHG = factor(inglesa_final$FTHG)

inglesa_final$FTAG = ifelse(inglesa_final$FTAG>=4, "4 o más", inglesa_final$FTAG)
inglesa_final$FTAG = factor(inglesa_final$FTAG)
```

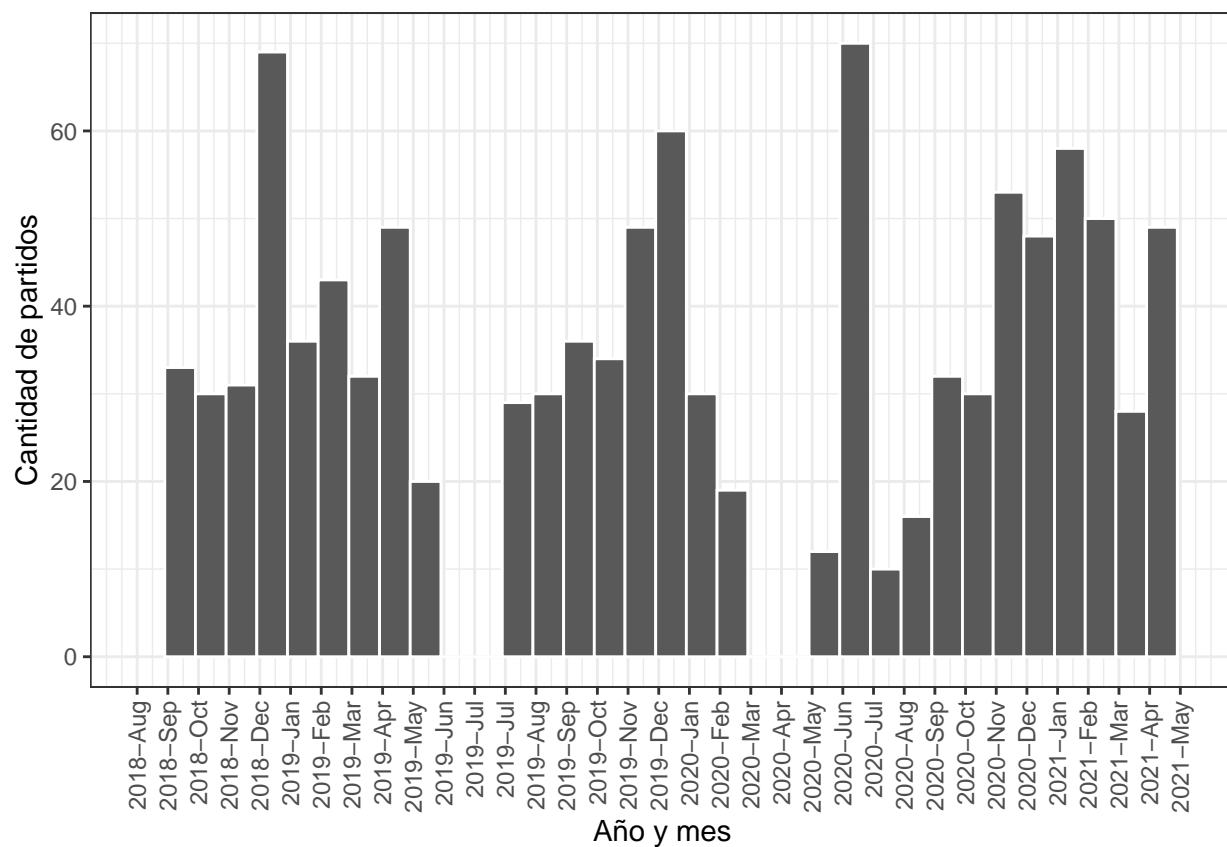
Resultados de la exploración de datos

Análisis Univariado

```
ggplot(inglesa_final, aes(x=Date)) + geom_histogram(binwidth=30, colour="white") + scale
```

Date

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



```
tabla_freq(inglesa_final$Date)
```

##	Categoría	Freq. Abs.	Freq. Rel.
## 1	2018-08-10	1	0.0008826125
## 2	2018-08-11	6	0.0052956752
## 3	2018-08-12	3	0.0026478376
## 4	2018-08-18	6	0.0052956752
## 5	2018-08-19	2	0.0017652251
## 6	2018-08-25	2	0.0017652251
## 7	2018-08-26	3	0.0026478376
## 8	2018-08-27	1	0.0008826125
## 9	2018-09-01	7	0.0061782877
## 10	2018-09-02	3	0.0026478376
## 11	2018-09-15	7	0.0061782877
## 12	2018-09-16	2	0.0017652251
## 13	2018-09-17	1	0.0008826125
## 14	2018-09-22	8	0.0070609003
## 15	2018-09-23	2	0.0017652251
## 16	2018-09-29	8	0.0070609003
## 17	2018-09-30	1	0.0008826125
## 18	2018-10-01	1	0.0008826125

## 19	2018-10-05	1	0.0008826125
## 20	2018-10-06	6	0.0052956752
## 21	2018-10-07	3	0.0026478376
## 22	2018-10-20	8	0.0070609003
## 23	2018-10-21	1	0.0008826125
## 24	2018-10-22	1	0.0008826125
## 25	2018-10-27	6	0.0052956752
## 26	2018-10-28	3	0.0026478376
## 27	2018-10-29	1	0.0008826125
## 28	2018-11-03	7	0.0061782877
## 29	2018-11-04	2	0.0017652251
## 30	2018-11-05	1	0.0008826125
## 31	2018-11-10	6	0.0052956752
## 32	2018-11-11	4	0.0035304501
## 33	2018-11-24	7	0.0061782877
## 34	2018-11-25	2	0.0017652251
## 35	2018-11-26	1	0.0008826125
## 36	2018-11-30	1	0.0008826125
## 37	2018-12-01	6	0.0052956752
## 38	2018-12-02	3	0.0026478376
## 39	2018-12-04	4	0.0035304501
## 40	2018-12-05	6	0.0052956752
## 41	2018-12-08	8	0.0070609003
## 42	2018-12-09	1	0.0008826125
## 43	2018-12-10	1	0.0008826125
## 44	2018-12-15	7	0.0061782877
## 45	2018-12-16	3	0.0026478376
## 46	2018-12-21	1	0.0008826125
## 47	2018-12-22	8	0.0070609003
## 48	2018-12-23	1	0.0008826125
## 49	2018-12-26	9	0.0079435128
## 50	2018-12-27	1	0.0008826125
## 51	2018-12-29	6	0.0052956752
## 52	2018-12-30	4	0.0035304501
## 53	2019-01-01	3	0.0026478376
## 54	2019-01-02	6	0.0052956752
## 55	2019-01-03	1	0.0008826125
## 56	2019-01-12	7	0.0061782877
## 57	2019-01-13	2	0.0017652251
## 58	2019-01-14	1	0.0008826125
## 59	2019-01-19	8	0.0070609003
## 60	2019-01-20	2	0.0017652251
## 61	2019-01-29	6	0.0052956752
## 62	2019-01-30	4	0.0035304501
## 63	2019-02-02	7	0.0061782877

## 64	2019-02-03	2	0.0017652251
## 65	2019-02-04	1	0.0008826125
## 66	2019-02-06	1	0.0008826125
## 67	2019-02-09	7	0.0061782877
## 68	2019-02-10	2	0.0017652251
## 69	2019-02-11	1	0.0008826125
## 70	2019-02-22	2	0.0017652251
## 71	2019-02-23	4	0.0035304501
## 72	2019-02-24	2	0.0017652251
## 73	2019-02-26	4	0.0035304501
## 74	2019-02-27	6	0.0052956752
## 75	2019-03-02	7	0.0061782877
## 76	2019-03-03	3	0.0026478376
## 77	2019-03-09	7	0.0061782877
## 78	2019-03-10	3	0.0026478376
## 79	2019-03-16	3	0.0026478376
## 80	2019-03-17	2	0.0017652251
## 81	2019-03-30	7	0.0061782877
## 82	2019-03-31	2	0.0017652251
## 83	2019-04-01	1	0.0008826125
## 84	2019-04-02	2	0.0017652251
## 85	2019-04-03	3	0.0026478376
## 86	2019-04-05	1	0.0008826125
## 87	2019-04-06	3	0.0026478376
## 88	2019-04-07	1	0.0008826125
## 89	2019-04-08	1	0.0008826125
## 90	2019-04-12	1	0.0008826125
## 91	2019-04-13	6	0.0052956752
## 92	2019-04-14	2	0.0017652251
## 93	2019-04-15	1	0.0008826125
## 94	2019-04-16	1	0.0008826125
## 95	2019-04-20	6	0.0052956752
## 96	2019-04-21	3	0.0026478376
## 97	2019-04-22	1	0.0008826125
## 98	2019-04-23	2	0.0017652251
## 99	2019-04-24	2	0.0017652251
## 100	2019-04-26	1	0.0008826125
## 101	2019-04-27	6	0.0052956752
## 102	2019-04-28	3	0.0026478376
## 103	2019-05-03	1	0.0008826125
## 104	2019-05-04	5	0.0044130627
## 105	2019-05-05	3	0.0026478376
## 106	2019-05-06	1	0.0008826125
## 107	2019-05-12	10	0.0088261253
## 108	2019-08-09	1	0.0008826125

## 109	2019-08-10	6	0.0052956752
## 110	2019-08-11	2	0.0017652251
## 111	2019-08-17	7	0.0061782877
## 112	2019-08-18	2	0.0017652251
## 113	2019-08-19	1	0.0008826125
## 114	2019-08-23	1	0.0008826125
## 115	2019-08-24	6	0.0052956752
## 116	2019-08-25	3	0.0026478376
## 117	2019-08-31	8	0.0070609003
## 118	2019-09-01	2	0.0017652251
## 119	2019-09-14	7	0.0061782877
## 120	2019-09-15	2	0.0017652251
## 121	2019-09-16	1	0.0008826125
## 122	2019-09-20	1	0.0008826125
## 123	2019-09-21	5	0.0044130627
## 124	2019-09-22	4	0.0035304501
## 125	2019-09-28	8	0.0070609003
## 126	2019-09-29	1	0.0008826125
## 127	2019-09-30	1	0.0008826125
## 128	2019-10-05	6	0.0052956752
## 129	2019-10-06	4	0.0035304501
## 130	2019-10-19	8	0.0070609003
## 131	2019-10-20	1	0.0008826125
## 132	2019-10-21	1	0.0008826125
## 133	2019-10-25	1	0.0008826125
## 134	2019-10-26	5	0.0044130627
## 135	2019-10-27	4	0.0035304501
## 136	2019-11-02	8	0.0070609003
## 137	2019-11-03	2	0.0017652251
## 138	2019-11-08	1	0.0008826125
## 139	2019-11-09	6	0.0052956752
## 140	2019-11-10	3	0.0026478376
## 141	2019-11-23	8	0.0070609003
## 142	2019-11-24	1	0.0008826125
## 143	2019-11-25	1	0.0008826125
## 144	2019-11-30	6	0.0052956752
## 145	2019-12-01	4	0.0035304501
## 146	2019-12-03	2	0.0017652251
## 147	2019-12-04	6	0.0052956752
## 148	2019-12-05	2	0.0017652251
## 149	2019-12-07	5	0.0044130627
## 150	2019-12-08	4	0.0035304501
## 151	2019-12-09	1	0.0008826125
## 152	2019-12-14	6	0.0052956752
## 153	2019-12-15	3	0.0026478376

## 154	2019-12-16	1	0.0008826125
## 155	2019-12-21	7	0.0061782877
## 156	2019-12-22	2	0.0017652251
## 157	2019-12-26	9	0.0079435128
## 158	2019-12-27	1	0.0008826125
## 159	2019-12-28	7	0.0061782877
## 160	2019-12-29	3	0.0026478376
## 161	2020-01-01	9	0.0079435128
## 162	2020-01-02	1	0.0008826125
## 163	2020-01-10	1	0.0008826125
## 164	2020-01-11	7	0.0061782877
## 165	2020-01-12	2	0.0017652251
## 166	2020-01-18	8	0.0070609003
## 167	2020-01-19	2	0.0017652251
## 168	2020-01-21	6	0.0052956752
## 169	2020-01-22	3	0.0026478376
## 170	2020-01-23	1	0.0008826125
## 171	2020-01-29	1	0.0008826125
## 172	2020-02-01	8	0.0070609003
## 173	2020-02-02	2	0.0017652251
## 174	2020-02-08	2	0.0017652251
## 175	2020-02-09	1	0.0008826125
## 176	2020-02-14	1	0.0008826125
## 177	2020-02-15	2	0.0017652251
## 178	2020-02-16	2	0.0017652251
## 179	2020-02-17	1	0.0008826125
## 180	2020-02-19	1	0.0008826125
## 181	2020-02-22	6	0.0052956752
## 182	2020-02-23	3	0.0026478376
## 183	2020-02-24	1	0.0008826125
## 184	2020-02-28	1	0.0008826125
## 185	2020-02-29	5	0.0044130627
## 186	2020-03-01	2	0.0017652251
## 187	2020-03-07	7	0.0061782877
## 188	2020-03-08	2	0.0017652251
## 189	2020-03-09	1	0.0008826125
## 190	2020-06-17	2	0.0017652251
## 191	2020-06-19	2	0.0017652251
## 192	2020-06-20	4	0.0035304501
## 193	2020-06-21	3	0.0026478376
## 194	2020-06-22	1	0.0008826125
## 195	2020-06-23	2	0.0017652251
## 196	2020-06-24	5	0.0044130627
## 197	2020-06-25	3	0.0026478376
## 198	2020-06-27	1	0.0008826125

## 199	2020-06-28	1	0.0008826125
## 200	2020-06-29	1	0.0008826125
## 201	2020-06-30	1	0.0008826125
## 202	2020-07-01	4	0.0035304501
## 203	2020-07-02	2	0.0017652251
## 204	2020-07-04	5	0.0044130627
## 205	2020-07-05	4	0.0035304501
## 206	2020-07-06	1	0.0008826125
## 207	2020-07-07	3	0.0026478376
## 208	2020-07-08	4	0.0035304501
## 209	2020-07-09	3	0.0026478376
## 210	2020-07-11	5	0.0044130627
## 211	2020-07-12	4	0.0035304501
## 212	2020-07-13	1	0.0008826125
## 213	2020-07-14	1	0.0008826125
## 214	2020-07-15	4	0.0035304501
## 215	2020-07-16	4	0.0035304501
## 216	2020-07-17	1	0.0008826125
## 217	2020-07-18	1	0.0008826125
## 218	2020-07-19	2	0.0017652251
## 219	2020-07-20	3	0.0026478376
## 220	2020-07-21	2	0.0017652251
## 221	2020-07-22	2	0.0017652251
## 222	2020-07-26	10	0.0088261253
## 223	2020-09-12	4	0.0035304501
## 224	2020-09-13	2	0.0017652251
## 225	2020-09-14	2	0.0017652251
## 226	2020-09-19	4	0.0035304501
## 227	2020-09-20	4	0.0035304501
## 228	2020-09-21	2	0.0017652251
## 229	2020-09-26	4	0.0035304501
## 230	2020-09-27	4	0.0035304501
## 231	2020-09-28	2	0.0017652251
## 232	2020-10-03	4	0.0035304501
## 233	2020-10-04	6	0.0052956752
## 234	2020-10-17	4	0.0035304501
## 235	2020-10-18	4	0.0035304501
## 236	2020-10-19	2	0.0017652251
## 237	2020-10-23	1	0.0008826125
## 238	2020-10-24	4	0.0035304501
## 239	2020-10-25	3	0.0026478376
## 240	2020-10-26	2	0.0017652251
## 241	2020-10-30	1	0.0008826125
## 242	2020-10-31	3	0.0026478376
## 243	2020-11-01	4	0.0035304501

## 244	2020-11-02	2	0.0017652251
## 245	2020-11-06	2	0.0017652251
## 246	2020-11-07	4	0.0035304501
## 247	2020-11-08	4	0.0035304501
## 248	2020-11-21	4	0.0035304501
## 249	2020-11-22	4	0.0035304501
## 250	2020-11-23	2	0.0017652251
## 251	2020-11-27	1	0.0008826125
## 252	2020-11-28	4	0.0035304501
## 253	2020-11-29	3	0.0026478376
## 254	2020-11-30	2	0.0017652251
## 255	2020-12-05	4	0.0035304501
## 256	2020-12-06	4	0.0035304501
## 257	2020-12-07	1	0.0008826125
## 258	2020-12-11	1	0.0008826125
## 259	2020-12-12	4	0.0035304501
## 260	2020-12-13	5	0.0044130627
## 261	2020-12-15	2	0.0017652251
## 262	2020-12-16	6	0.0052956752
## 263	2020-12-17	2	0.0017652251
## 264	2020-12-19	4	0.0035304501
## 265	2020-12-20	4	0.0035304501
## 266	2020-12-21	2	0.0017652251
## 267	2020-12-26	6	0.0052956752
## 268	2020-12-27	4	0.0035304501
## 269	2020-12-28	2	0.0017652251
## 270	2020-12-29	5	0.0044130627
## 271	2020-12-30	1	0.0008826125
## 272	2021-01-01	2	0.0017652251
## 273	2021-01-02	4	0.0035304501
## 274	2021-01-03	2	0.0017652251
## 275	2021-01-04	1	0.0008826125
## 276	2021-01-12	3	0.0026478376
## 277	2021-01-13	2	0.0017652251
## 278	2021-01-14	1	0.0008826125
## 279	2021-01-16	5	0.0044130627
## 280	2021-01-17	3	0.0026478376
## 281	2021-01-18	1	0.0008826125
## 282	2021-01-19	2	0.0017652251
## 283	2021-01-20	2	0.0017652251
## 284	2021-01-21	1	0.0008826125
## 285	2021-01-23	1	0.0008826125
## 286	2021-01-26	4	0.0035304501
## 287	2021-01-27	5	0.0044130627
## 288	2021-01-28	1	0.0008826125

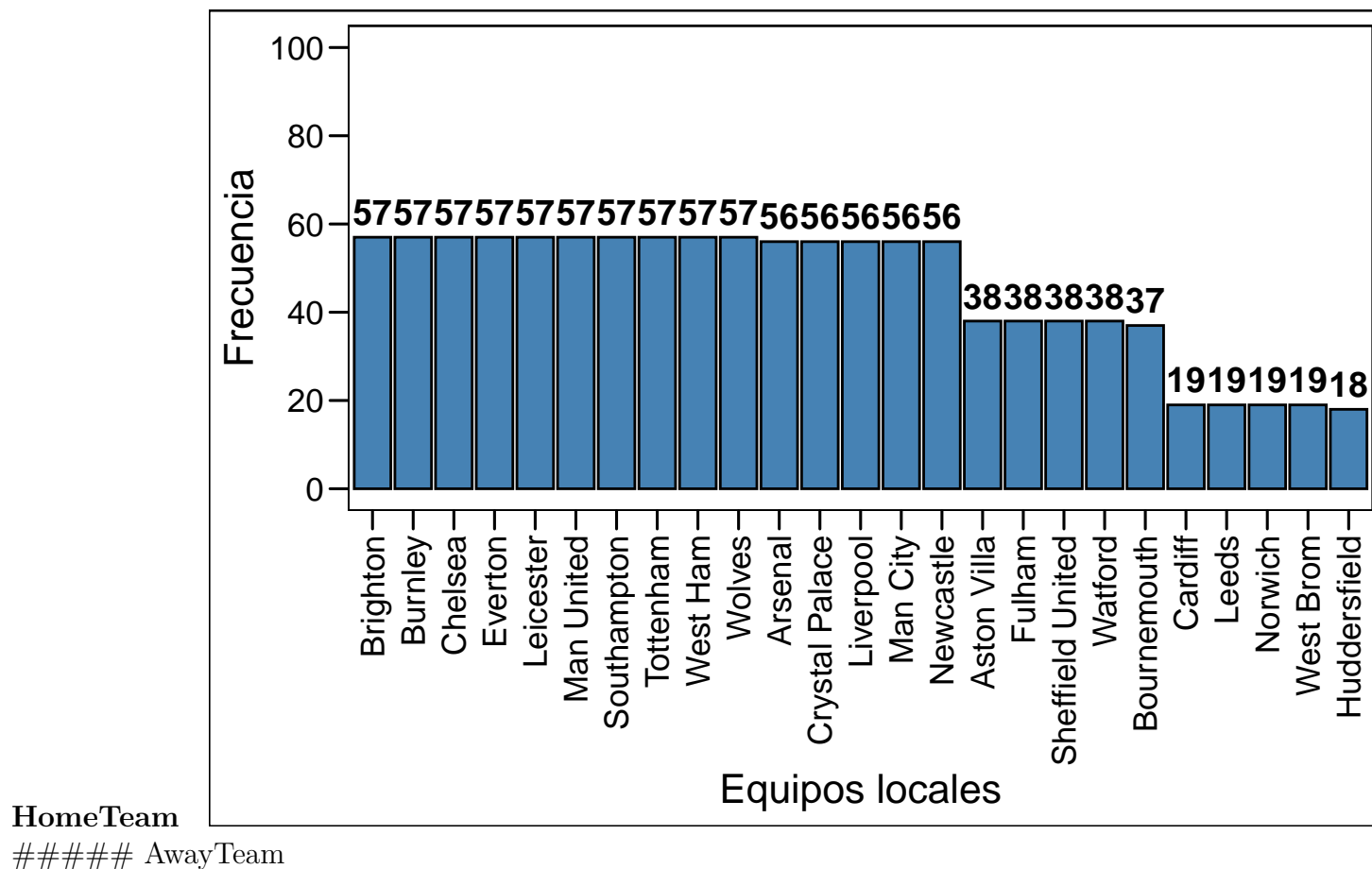
## 289	2021-01-30	6	0.0052956752
## 290	2021-01-31	4	0.0035304501
## 291	2021-02-02	4	0.0035304501
## 292	2021-02-03	5	0.0044130627
## 293	2021-02-04	1	0.0008826125
## 294	2021-02-06	5	0.0044130627
## 295	2021-02-07	4	0.0035304501
## 296	2021-02-08	1	0.0008826125
## 297	2021-02-13	4	0.0035304501
## 298	2021-02-14	4	0.0035304501
## 299	2021-02-15	2	0.0017652251
## 300	2021-02-17	2	0.0017652251
## 301	2021-02-19	1	0.0008826125
## 302	2021-02-20	4	0.0035304501
## 303	2021-02-21	4	0.0035304501
## 304	2021-02-22	1	0.0008826125
## 305	2021-02-23	1	0.0008826125
## 306	2021-02-27	4	0.0035304501
## 307	2021-02-28	5	0.0044130627
## 308	2021-03-01	1	0.0008826125
## 309	2021-03-02	1	0.0008826125
## 310	2021-03-03	3	0.0026478376
## 311	2021-03-04	3	0.0026478376
## 312	2021-03-06	4	0.0035304501
## 313	2021-03-07	4	0.0035304501
## 314	2021-03-08	2	0.0017652251
## 315	2021-03-10	1	0.0008826125
## 316	2021-03-12	1	0.0008826125
## 317	2021-03-13	4	0.0035304501
## 318	2021-03-14	4	0.0035304501
## 319	2021-03-15	1	0.0008826125
## 320	2021-03-19	1	0.0008826125
## 321	2021-03-20	1	0.0008826125
## 322	2021-03-21	2	0.0017652251
## 323	2021-04-03	4	0.0035304501
## 324	2021-04-04	4	0.0035304501
## 325	2021-04-05	2	0.0017652251
## 326	2021-04-09	1	0.0008826125
## 327	2021-04-10	3	0.0026478376
## 328	2021-04-11	4	0.0035304501
## 329	2021-04-12	2	0.0017652251
## 330	2021-04-16	1	0.0008826125
## 331	2021-04-17	2	0.0017652251
## 332	2021-04-18	2	0.0017652251
## 333	2021-04-19	1	0.0008826125

## 334	2021-04-20	1	0.0008826125
## 335	2021-04-21	2	0.0017652251
## 336	2021-04-22	1	0.0008826125
## 337	2021-04-23	1	0.0008826125
## 338	2021-04-24	3	0.0026478376
## 339	2021-04-25	3	0.0026478376
## 340	2021-04-26	1	0.0008826125
## 341	2021-04-30	1	0.0008826125
## 342	2021-05-01	4	0.0035304501
## 343	2021-05-02	2	0.0017652251
## 344	2021-05-03	2	0.0017652251
## 345	2021-05-07	1	0.0008826125
## 346	2021-05-08	4	0.0035304501
## 347	2021-05-09	4	0.0035304501
## 348	2021-05-10	1	0.0008826125
## 349	2021-05-11	2	0.0017652251
## 350	2021-05-12	1	0.0008826125
## 351	2021-05-13	2	0.0017652251
## 352	2021-05-14	1	0.0008826125
## 353	2021-05-15	3	0.0026478376
## 354	2021-05-16	4	0.0035304501
## 355	2021-05-18	4	0.0035304501
## 356	2021-05-19	6	0.0052956752
## 357	2021-05-23	10	0.0088261253
## 358	Total	1133	1.0000000000

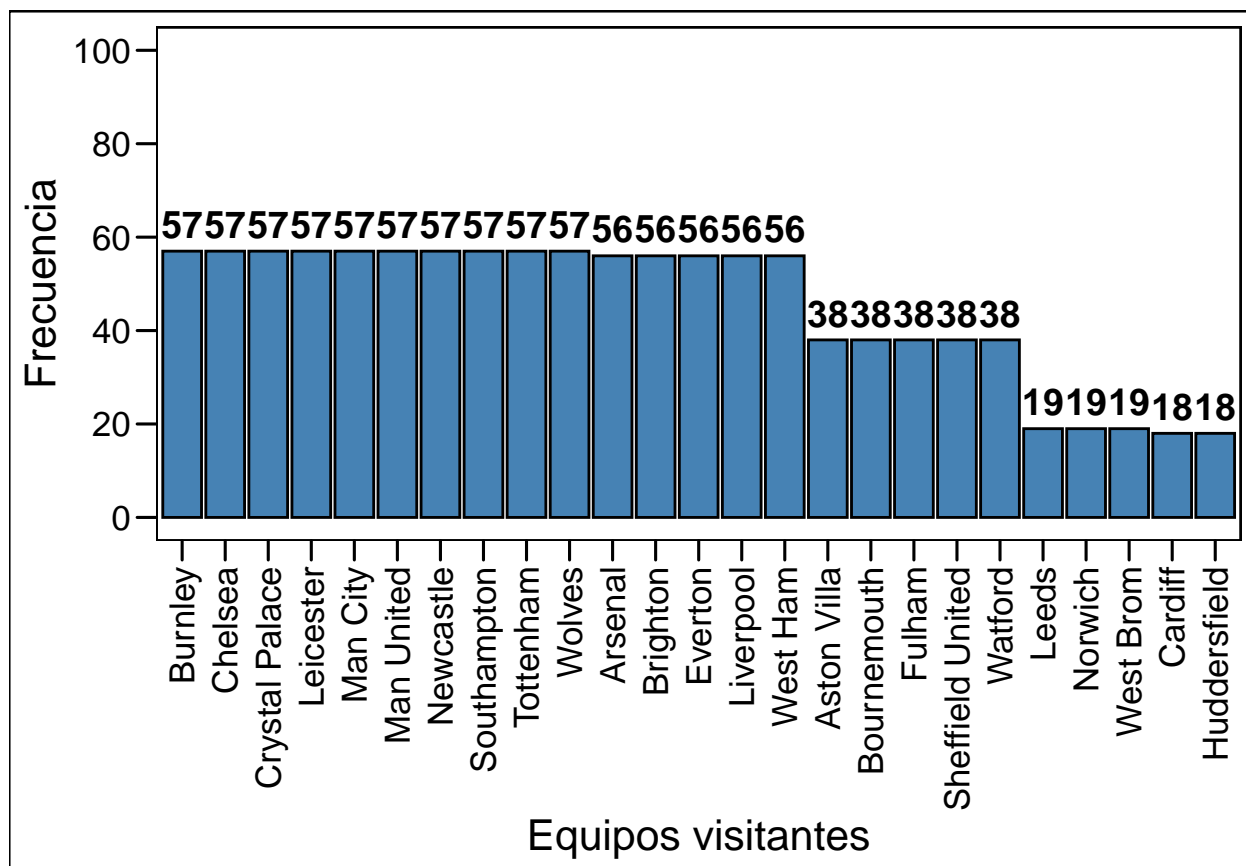
```

TABLA_HOMETEAM=tabla_freq(inglesa_final$HomeTeam,0)
#FRECUENCIA ABSOLUTA
ggplot(TABLA_HOMETEAM, aes(x=reorder(Categoría, -`Freq. Abs.`), y=`Freq. Abs.`))+
  geom_bar(stat = "identity", fill="steelblue", color="black")+
  labs(x="Equipos locales",y="Frecuencia")+
  scale_y_continuous(limits = c(0,100), breaks = seq(0,100,20))+
  geom_text(aes(x=Categoría, y=`Freq. Abs.`), label=`Freq. Abs.`),vjust=-0.5, fontface =
  theme_base() + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

```



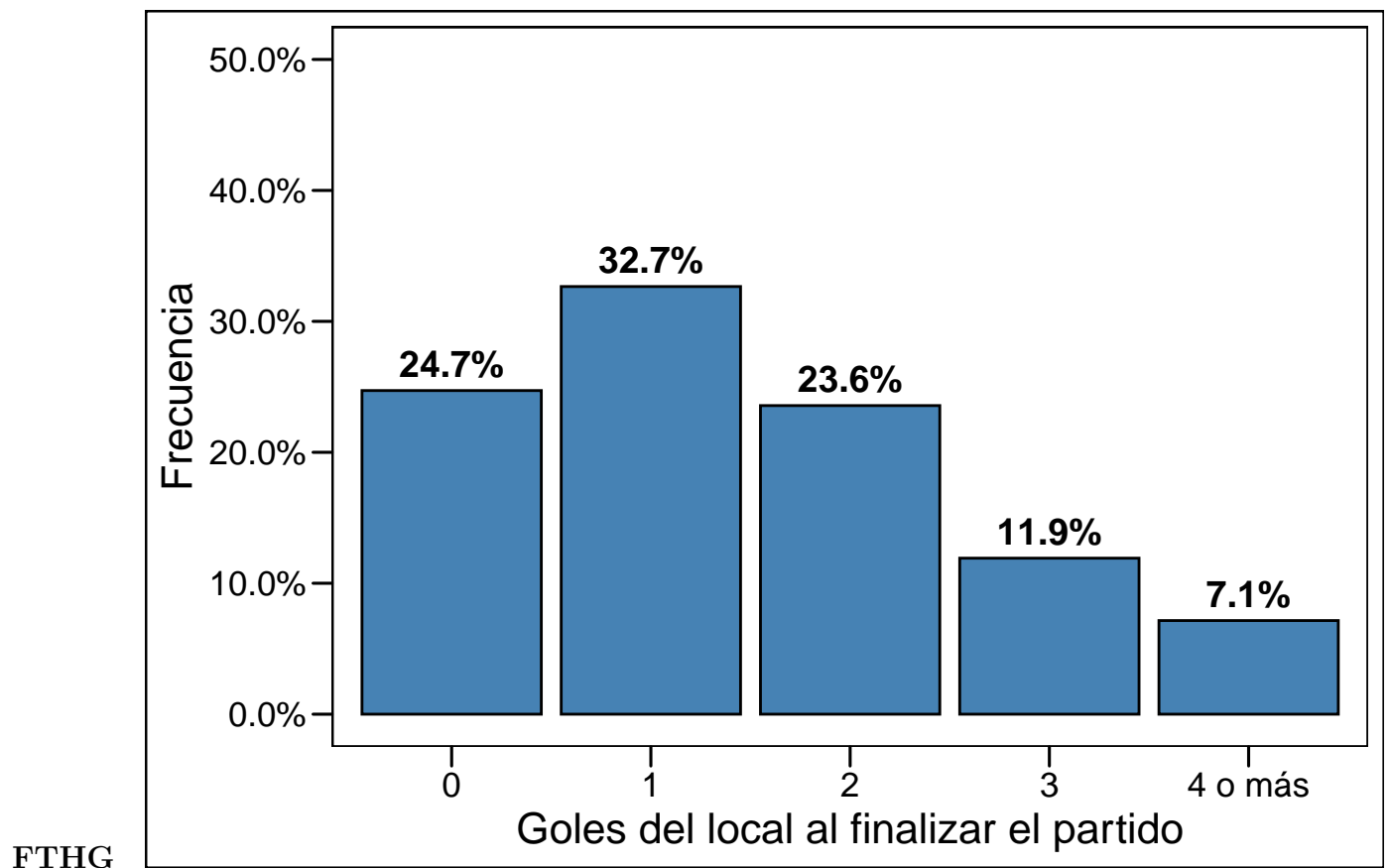
```
TABLA_AWAYTEAM=tabla_freq(inglesa_final$AwayTeam,0)
#FRECUENCIA ABSOLUTA
ggplot(TABLA_AWAYTEAM, aes(x=reorder(Categoría, -`Freq. Abs.`), y=`Freq. Abs.`))+
  geom_bar(stat = "identity", fill="steelblue", color="black")+
  labs(x="Equipos visitantes",y="Frecuencia")+
  scale_y_continuous(limits = c(0,100), breaks = seq(0,100,20))+
  geom_text(aes(x=Categoría, y=`Freq. Abs.`), label=`Freq. Abs.`),vjust=-0.5, fontface =
  theme_base() + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



```
TABLA_FTHG=tabla_freq(inglesa_final$FTHG,0)
```

```
#FRECUENCIA RELATIVA
```

```
ggplot(TABLA_FTHG, aes(x=Categoría, y=`Freq. Rel.`))+
  geom_bar(stat = "identity", fill="steelblue", color="black")+
  labs(x="Goles del local al finalizar el partido",y="Frecuencia")+
  scale_y_continuous(limits = c(0,ifelse(max(TABLA_FTHG$`Freq. Rel.`)>0.5,1,0.5)), break
  geom_text(aes(x=Categoría, y=`Freq. Rel.` , label=percent(`Freq. Rel.`)),vjust=-0.5, fo
  theme_base()
```

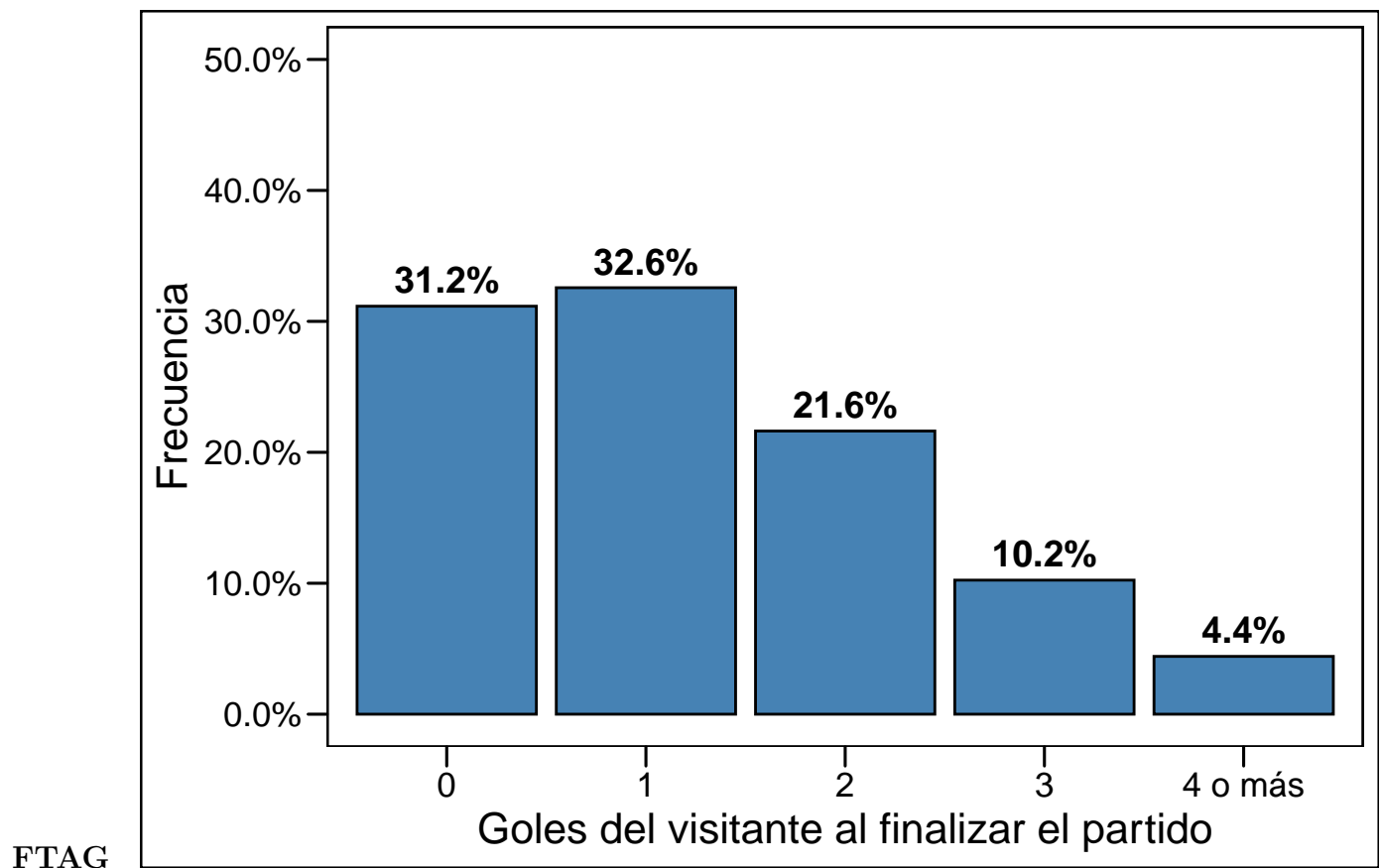


La mayoría de los equipos locales hicieron 1 gol al finalizar el partido y la minoría hizo 4 o más goles.

```
TABLA_FTAG=tabla_freq(inglesa_final$FTAG,0)
```

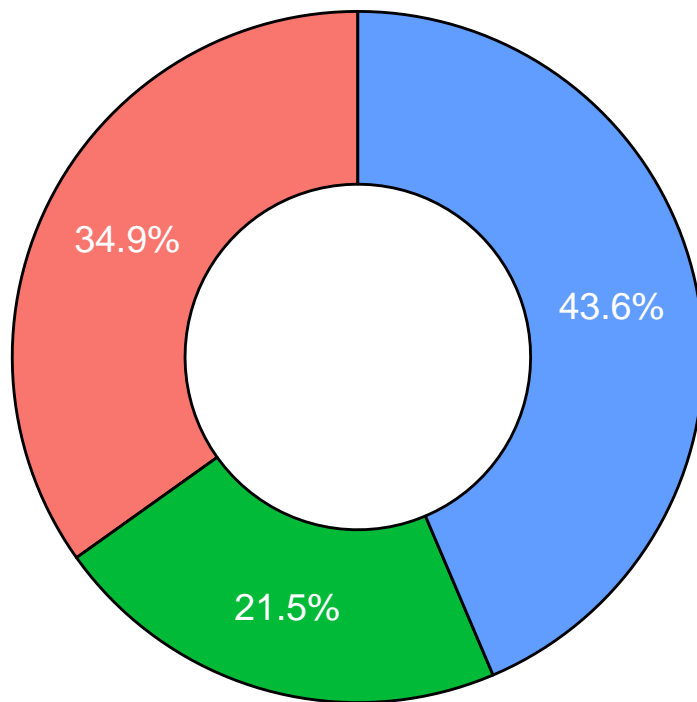
```
#FRECUENCIA RELATIVA
```

```
ggplot(TABLA_FTAG, aes(x=Categoría, y=`Freq. Rel.`))+
  geom_bar(stat = "identity", fill="steelblue", color="black")+
  labs(x="Goles del visitante al finalizar el partido",y="Frecuencia")+
  scale_y_continuous(limits = c(0,ifelse(max(TABLA_FTAG$`Freq. Rel.`)>0.5,1,0.5)), break
  geom_text(aes(x=Categoría, y=`Freq. Rel.`), label=percent(`Freq. Rel.`),vjust=-0.5, fo
  theme_base()
```

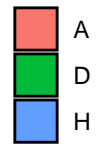


La mayoría de los equipos visitantes hicieron 1 gol al finalizar el partido y la minoría hizo 4 o más goles.

```
TABLA_FTR=tabla_freq(inglesa_final$FTR,0)
#FRECUENCIA ABSOLUTA
ggplot(TABLA_FTR, aes(x=2, y=`Freq. Rel.`, fill=Categoría)) +
  geom_bar(stat="identity", width=1, color="black") +
  coord_polar("y", start=0)+
  labs(fill="Marcador final del partido")+
  theme_void()+
  geom_text(aes(label = percent(`Freq. Rel.`)),position = position_stack(vjust = 0.5),color="black",size=12) +
  xlim(0.5, 2.5)
```



Marcador final del partido



FTR