NFL 1st and Future – Analytics

FEATURE IMPORTANCE, FEATURE INTERACTION, FACTOR ANALYSIS AND RECURSIVE FEATURE ELIMINATION

ABSTRACT

This review tries to examine the effects of factors including playing surfaces on player movements and by extension, their concurrence with lower extremity injuries.

The analysis uncovers specific features and feature interactions which combined with player movement present ideal conditions for risk of injury.

The analysis concentrates on game conditions instead of play conditions, as a quarter of the injury data has no identifying play key. The choice to concentrate on game conditions shows better predictive quality across different models than play conditions.

Our analysis discovers that the specific set of factors that influence lower extremity injuries are mutually exclusive to types of injuries.

Our analysis also reinforces the idea that synthetic playing surfaces play a major role in occurrence of these injuries.

The analysis does not discredit player movements with contributing to injuries, however it proposes that the information given is inadequate as we cannot confirm at what time point in the duration of the play that the injury occurred. Therefore, our representation of speed, directional changes, acceleration and distance produced no predictive qualities.

The analysis is bi-directional: where features are found to promote risk, we look for ways to reduce or eliminate those factors. Where features are found to have low correlation to risk, we look for ways to magnify or encourage their occurrence.

INTRODUCTION

Our analysis examined eight crowd-sourced hypotheses:

1. **Special teams**: That players on special teams were more prone to injuries
2. **Injuries during practice/preseason**: That players were prone to more injuries during practice as there are more practice sessions that regular season games
3. **Backups more prone to injury:** That substitutes were more prone to injuries
4. **Coaching decisions:** That coaching decisions have an influence on injuries
5. **Coming back from injury/% healthy:** That players were more prone to injury immediately after coming back from injury

6. **Injured in cold temperatures/more start/stops:** That 'slower' games have a higher prevalence of injury
7. **Same injury, more severe recovery time on synthetic:** That synthetic turf not only had a higher prevalence of injuries but a higher prevalence of more severe injuries
8. **Less injuries during wet weather:** That the effect of natural versus synthetic turf is eliminated during games with rain or snow.
9. **Games with multiple injuries:** That there are games with perfect conditions to cause multiple injuries.

DATA SET AND FEATURE ENGINEERING

Characteristics of the dataset included InjuryRecord missing 28 PlayKeys, PlayerTrackData missing 74,526,875 event values and PlayList missing 16910 StadiumType, 18691 Weather and 367 PlayType values. This is consequential because tree models have a selection bias against columns with missing values and imputed values might skew the analysis. We note the columns as they are likely not to depended on for analysis.

First step to handling the data was to free up memory by downgrading the datatypes for all data to make it easier to fit into memory.

All the datasets were merged by common primary keys and sliced for the row representing the termination of the play (where time on the play is maximum).

Fig 1: PlayList

Feature	*Definition*
IsWet	Is it raining or snowing? [0,1]
IsSunny	Is it sunny or clear? [0,1]
IsCloudy	Is it cloudy or overcast? [0,1]
IsSnow	Is it snowing? [0,1]
IsControlled	Is temperature controlled? [0,1]
IsDomeOpen	Is dome/bowl with roof open? [0,1]
IsDomeClosed	Is dome/bowl with roof closed? [0,1]
IsIndoor	Is played indoor? [0,1]
IsOutdoor	Is played outdoor? [0,1]
IsStadiumUnkown	Conditions unknown
PlayerDay0	Sequences player day participation in a 116-day season
Preseason	Signify [0,1] games played prior to the regular season
Temperature0	Binned temperature feature in 10-degree bins from zero
Weather0	Bucketing weather into snow, rain, clear, cloudy, unknown
StaduimType0	Bucketing stadiumtype into indoor, outdoor, dome open, dome closed

Fig 2: InjuryRecord

Feature	*Definition*
BodyPart	String for which bodypart was injured

injuryKnee	Binary [0,1] for knee injury
injuryAnkle	Binary [0,1] for ankle injury
injuryHeel	Binary [0,1] for heel injury
injuryToes	Binary [0,1] for toe injury
injuryFoot	Binary [0,1] for foot injury
RecoveryTime	Converts the one-hot encoded days for recovery into a numerical feature.

Fig 3: PlayerTrackData

Feature	*Definition*
v_horizontal	Calculates angular speed on x-axis
v_vertical	Calculates angular speed on y-axis
x_vdiff_f1delta	Change in x over change in time
x_vdiff_adiff_f2delta	Change in change in x over change in time
y_vdiff_f1delta	Change in y over change in time
y_vdiff_adiff_f2delta	Change in change in y over change in time
dis_vdiff_f1delta	Change in dis over change in time
dis_vdiff_adiff_f2delta	Change in change in dis over change in time
dir_vdiff_f1delta	Change in dir over change in time
dir_vdiff_adiff_f2delta	Change in change in dir over change in time
o_vdiff_f1delta	Change in o over change in time
o_vdiff_adiff_f2delta	Change in change in o over change in time
s_vdiff_f1delta	Change in s over change in time
s_vdiff_adiff_f2delta	Change in change in s over change in time
x_vdiff_f1delta_max	Quantify extremes above group max
x_vdiff_adiff_f2delta_max	Quantify extremes above group max
y_vdiff_f1delta_max	Quantify extremes above group max
y_vdiff_adiff_f2delta_max	Quantify extremes above group max
dis_vdiff_f1delta_max	Quantify extremes above group max
dis_vdiff_adiff_f2delta_max	Quantify extremes above group max
dir_vdiff_f1delta_max	Quantify extremes above group max
dir_vdiff_adiff_f2delta_max	Quantify extremes above group max
o_vdiff_f1delta_max	Quantify extremes above group max
o_vdiff_adiff_f2delta_max	Quantify extremes above group max
s_vdiff_f1delta_max	Quantify extremes above group max
s_vdiff_adiff_f2delta_max	Quantify extremes above group max
x_vdiff_f1delta_std	Quantify extremes beyond group std
x_vdiff_adiff_f2delta_std	Quantify extremes beyond group std
y_vdiff_f1delta_std	Quantify extremes beyond group std
y_vdiff_adiff_f2delta_std	Quantify extremes beyond group std
dis_vdiff_f1delta_std	Quantify extremes beyond group std
dis_vdiff_adiff_f2delta_std	Quantify extremes beyond group std
dir_vdiff_f1delta_std	Quantify extremes beyond group std

dir_vdiff_adiff_f2delta_std	Quantify extremes beyond group std
o_vdiff_f1delta_std	Quantify extremes beyond group std
o_vdiff_adiff_f2delta_std	Quantify extremes beyond group std
s_vdiff_f1delta_std	Quantify extremes beyond group std
s_vdiff_adiff_f2delta_std	Quantify extremes beyond group std

**Mean and standard deviation for speed, distance, acceleration features are contained in the function called createstat() but having poor predictive quality, they are commented out in the notebook.*

The dataset was then split into three sets:

1. Training dataset (80%)
2. Test dataset (20%)
    a. Evaluation dataset (10%)
    b. Inference and Scoring (10%)

MODELS AND IMPLEMENTATION

Our binary classification algorithm:

1. We create two datasets: baseline data and the feature engineered dataset.
2. Select individual injury to be analyzed (Features affecting different injuries are exclusive – see Data Augmentation section)
3. Split dataset into training, evaluation and prediction set.
4. Weighted parameters are set for each class to address class imbalance in the dataset
5. Where available, L1 and L2 regularization is set for feature selection
6. Run prediction model
7. Use fitted model to predict on prediction dataset
1. Calculate scoring on prediction dataset labels vs predictions (precision, recall, fscore)
8. Calculate confusion matrix
9. Calculate feature importance/feature interaction

Three models were selected, each with their own strengths and weaknesses in binary classification.

Catboost was selected as it does not need one-hot encoded categorical features, which allows for easy feature engineering by mixing numerical features and string features. Catboost allows for the feature importance, feature interaction importance, hyperparameter grid search and object importance. This model was used to rank feature importance as well as calculate what pairwise feature interactions had the best predictive qualities.

LightGBM was selected as a control for the CatBoost model as it implements the same gradient boosting algorithm and is faster. This model was used to countercheck catboost feature importance and used for one step recursive feature elimination.

Random Forest was selected as it implements a different algorithm and therefore can be used as a comparison to CatBoost/LightGBM to triangulate the answer as well as countercheck. It was primarily used for feature importance.

DATA AUGMENTATION

The idea was to augment the small injury dataset by duplicating knee and ankle injuries data and reversing the injury label. This was based off Player with PlayerKey 47307 who managed to injure his knee and ankle on the same play in the same game. Since the conditions that led to the double injury were similar, the assumption was that conditions affecting knee and ankle injuries are common.

However, running a binary classification on the augmented dataset produced abnormally high log-loss scores i.e. 0.35 while non-augmented data achieved log-loss scores of 0.05. This proves that these two injuries had mutually exclusive sets of features/conditions affecting their occurrence.

Therefore, any analysis should consider all the injuries individually.

Fig 4:

	PlayerKey	**GameID**	**PlayKey**	**BodyPart**	**Surface**	**DM_M1**	**DM_M7**	**DM_M28**	**DM_M42**	**Recovery Time**
86	47307	47307-10	47307-10-18	Knee	Synthetic	1	1	0	0	7
87	47307	47307-10	47307-10-18	Ankle	Synthetic	1	1	0	0	7

**Data augmentation contained in the function called daugment() but having poor predictive quality, they are commented out in the notebook.*

ANALYSIS

1. INJURIES IN GENERAL

Factor Analysis:

Factor Analysis can be used as a form of exploratory feature analysis by dimension reduction.

Factor analysis was done to discover hidden relationships between features by extracting maximum common variance. After calculating eigen values, we create 13 factors which explain 0.6867 of the total variances. Even after calculating for 46 factors, the total explained variance was approximately 0.6521.

Fig 5:

```
fa.get_factor_variance()[2]
```

```
array([0.16088685, 0.24137573, 0.30582962, 0.35561795, 0.40022888,
       0.44186521, 0.48016781, 0.51374535, 0.54594585, 0.57409343,
       0.59862159, 0.61997265, 0.63967396, 0.65211681])
```

We then calculate the features that had the highest loadings on each of the factors and assume that these factors are linearly related to a smaller number of unobservable factors. The three top features for each loading is shown below:

Factor 0: dis_vdiff_adiff_f2delta, s_vdiff_adiff_f2delta , s_vdiff_f1delta (Cum. Var 0.1608)

Factor 1: v_horizontal , v_vertical , o_vdiff_adiff_f2delta (Cum. Var 0.2414)

Factor 2: IsIndoor , FieldType , IsDomeOpen (Cum. Var 0.3058)

Factor 3: Position, PositionGroup , RosterPosition (Cum. Var 0.3556)

Factor 4: x_vdiff_adiff_f2delta, x_vdiff_f1delta , y_vdiff_f1delta (Cum. Var 0.4002)

Factor 5: IsSunny , Weather , Temperature0 (Cum. Var 0.4419)

Factor 6: StadiumType , IsOutdoor , IsIndoor (Cum. Var 0.4802)

Factor 7: PlayerDay0, PlayerGame , IsSnow (Cum. Var 0.5137)

Factor 8: PlayType , PlayerGamePlay , PlayerGame (Cum. Var 0.5459)

Factor 9: IsWet , IsSnow , Weather (Cum. Var 0.5741)

Factor 10: s, dis, s_vdiff_adiff_f2delta (Cum. Var 0.5986)

Factor 11: dir , IsDomeOpen , StadiumType (Cum. Var 0.6200)

Factor 12: IsDomeOpen , StadiumType , Temperature0 (Cum. Var 0.6397)

Factor 13: y, IsDomeOpen , IsStadiumUnknown (Cum. Var 0.6521)

Fig 6:

```
0
                Features        0
13  dis_vdiff_adiff_f2delta  0.950547
19    s_vdiff_adiff_f2delta  0.950293
18          s_vdiff_f1delta  0.949956
1
                Features        1
6            v_horizontal  0.939741
7              v_vertical  0.932126
17  o_vdiff_adiff_f2delta  0.461854
2
      Features        2
37    IsIndoor  0.954627
23   FieldType  0.580196
35  IsDomeOpen  0.329771
3
         Features        3
27       Position  0.852509
28   PositionGroup  0.830790
20  RosterPosition  0.819823
4
                Features        4
9   x_vdiff_adiff_f2delta  0.958368
8         x_vdiff_f1delta  0.958299
10        y_vdiff_f1delta  0.149572
5
        Features        5
30       IsSunny  0.898577
24       Weather  0.359466
41  Temperature0  0.095524
6
       Features        6
22  StadiumType  0.823402
38    IsOutdoor  0.495025
37     IsIndoor  0.115312
7
      Features        7
39  PlayerDay0  1.012542
21  PlayerGame  0.333957
31      IsSnow  0.083536
8
         Features        8
25       PlayType  0.094385
26  PlayerGamePlay  0.032159
21      PlayerGame  0.024949
9
   Features        9
29    IsWet  0.990332
31   IsSnow  0.334073
24  Weather  0.107890
10
                Features       10
5                      s  0.833182
3                    dis  0.808984
19  s_vdiff_adiff_f2delta  0.195796
11
       Features       11
2           dir  0.065804
35   IsDomeOpen  0.065581
22  StadiumType  0.065456
12
        Features       12
35    IsDomeOpen  0.635640
22   StadiumType  0.517361
41  Temperature0  0.333144
13
           Features       13
1                 y  0.047642
35        IsDomeOpen  0.011053
34  IsStadiumUnknown  0.008529
```

Fig 7: In the figure, we see that synthetic turf injuries outnumber natural turf injuries in ankles and toes and the reverse is true in feet. Heel injuries only occurred on natural turf and knee injuries were equivalent.

```
dxp.aggplot(agg='BodyPart', data=injuries, hue='Surface',normalize='BodyPart')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f2559475be0>
```

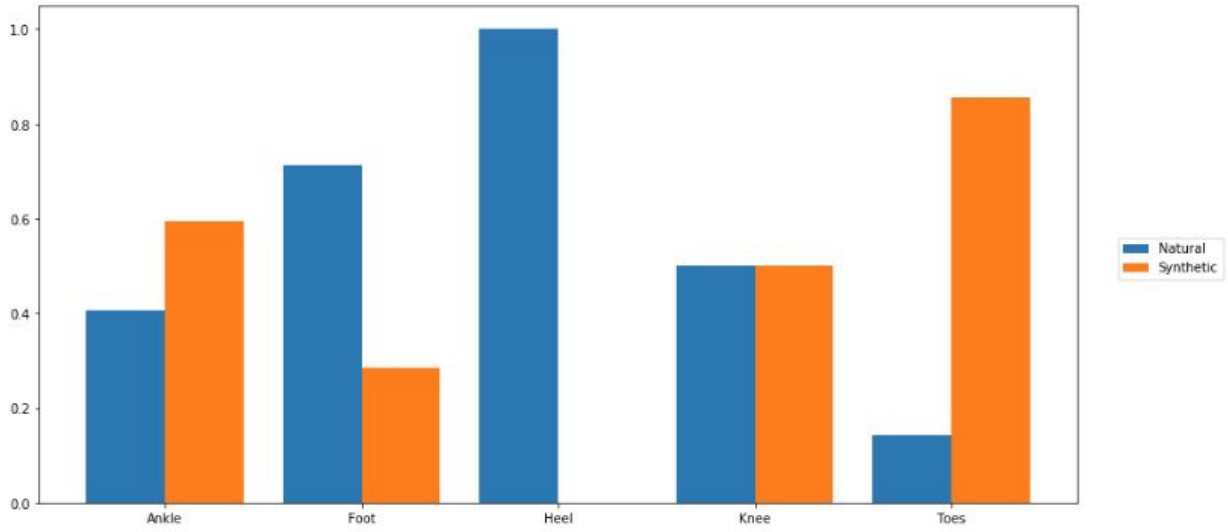


Fig 8: Across all injuries, knees made the majority followed by ankles.

```
dxp.aggplot(agg='BodyPart', data=injuries, hue='Surface',normalize='all')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f25593e2390>
```

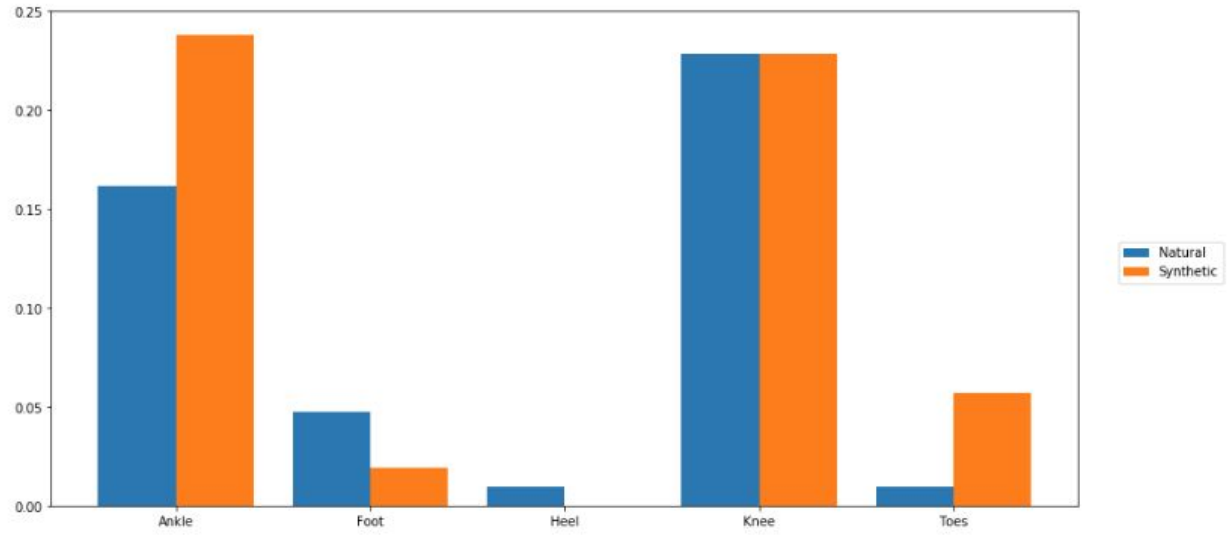


Fig 9: When normalized by surface type, knee injuries are the majority on natural turf while ankles are the majority on synthetic turf.

```
dxp.aggplot(agg='BodyPart', data=injuries, hue='Surface',normalize='Surface')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f2559353550>
```

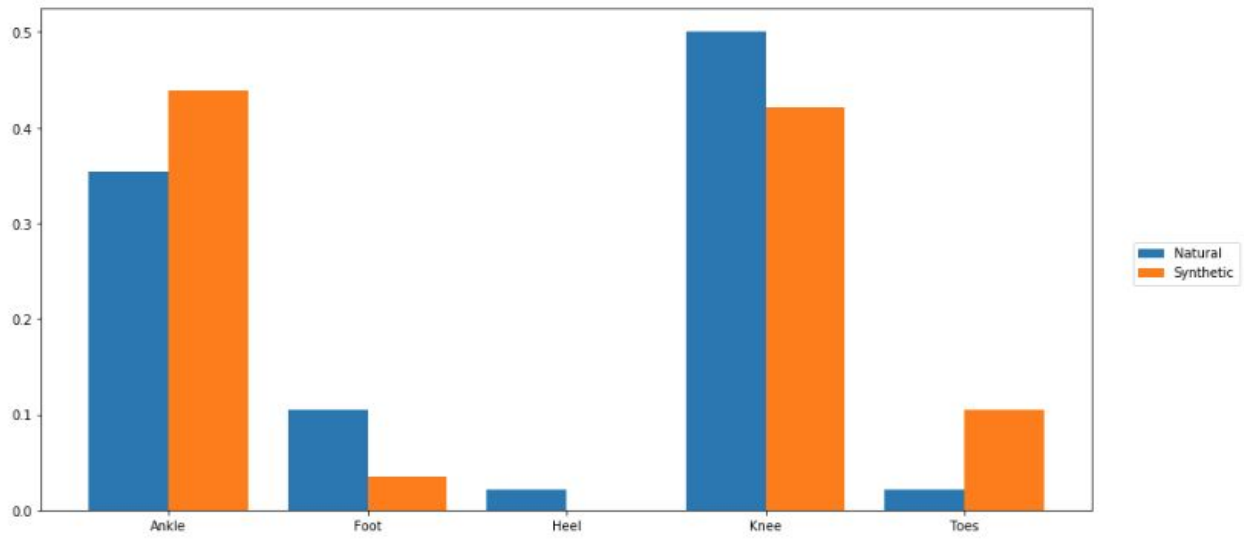


Fig 10: Outdoor stadiums are heavily represented in the injury data for knees, feet and ankles.

```
if (choice=='Knee') or (choice=='Foot') or (choice=='Ankle'):
    dxp.aggplot(agg='BodyPart', data=injuriesplays0, hue='StadiumType0',normalize='BodyPart')
```

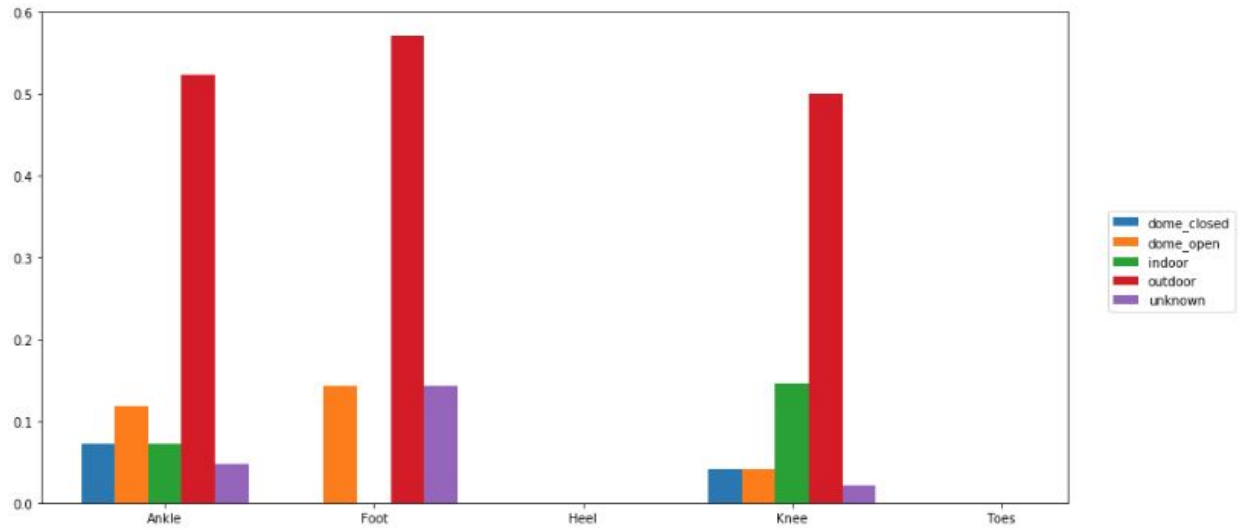


Fig 11: Across all injuries, outdoor stadiums are heavily represented.

```
if (choice=='Knee') or (choice=='Foot') or (choice=='Ankle'):
    dxp.aggplot(agg='BodyPart', data=injuriesplays0, hue='StadiumType0',normalize='all')
```

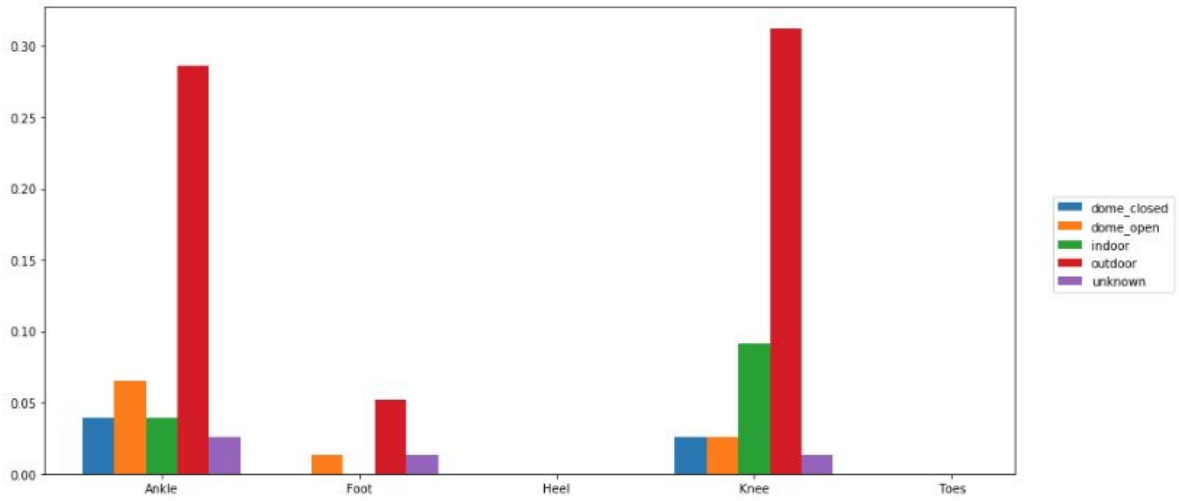


Fig 12: When normalized by stadiumtype, knee injuries have a higher prevalence in indoor stadiums and outdoor stadiums while ankle injuries in open dome stadiums.

```
if (choice=='Knee') or (choice=='Foot') or (choice=='Ankle'):
    dxp.aggplot(agg='BodyPart', data=injuriesplays0, hue='StadiumType0',normalize='StadiumTyp
e0')
```

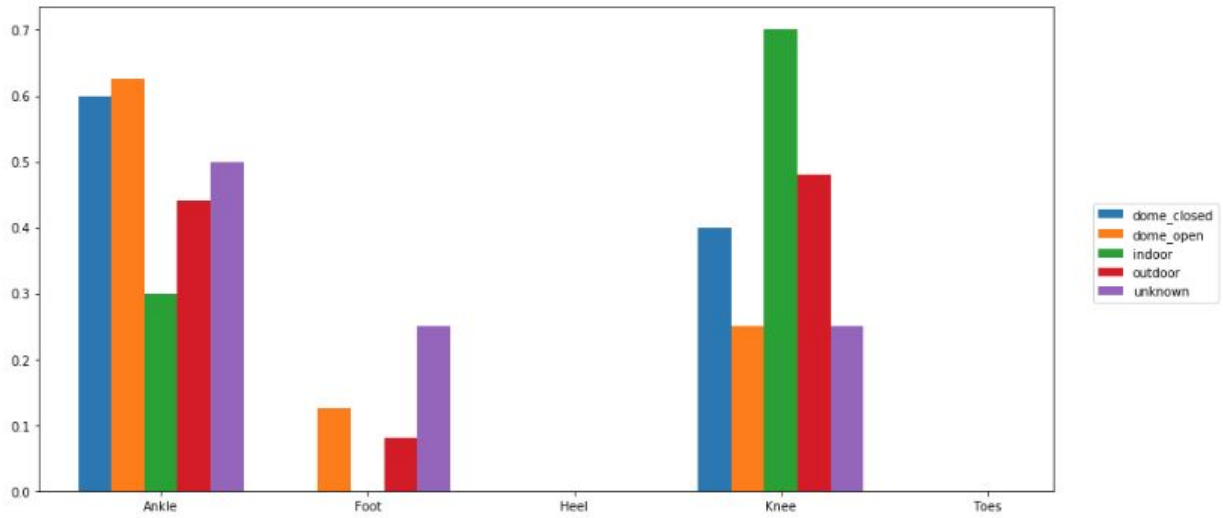


Fig 13: Knees and ankles afflict linebackers and wide receivers more while feet injuries afflict offensive linemen.

```
if (choice=='Knee') or (choice=='Foot') or (choice=='Ankle'):
    dxp.aggplot(agg='BodyPart', data=injuriesplays0, hue='RosterPosition',normalize='BodyPar
t')
```

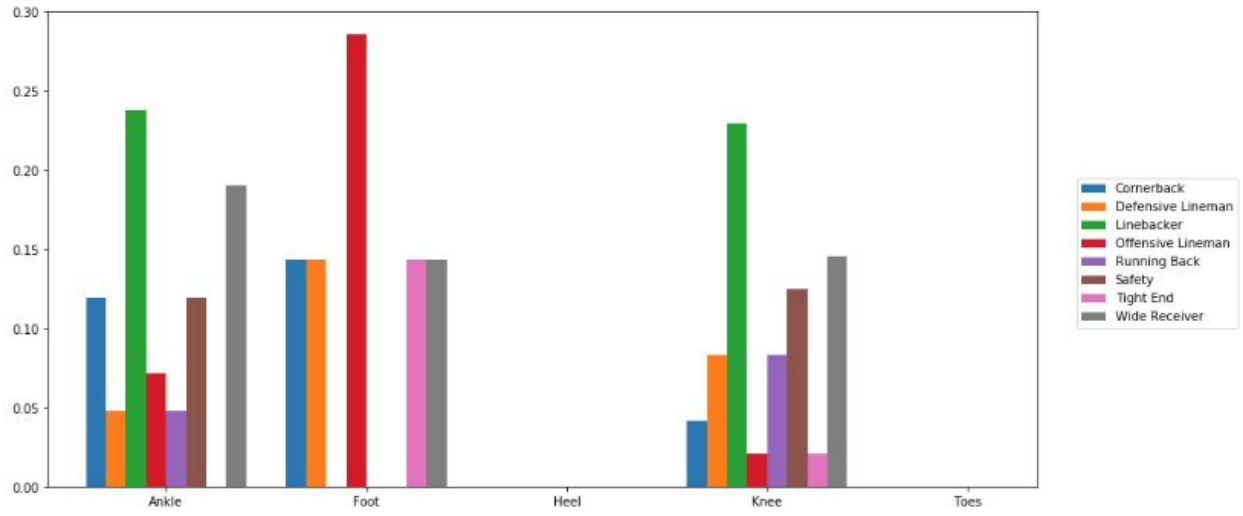


Fig 14: Across all injuries, linebackers and wide receivers were the groups most highly afflicted.

```
if (choice=='Knee') or (choice=='Foot') or (choice=='Ankle'):
    dxp.aggplot(agg='BodyPart', data=injuriesplays0, hue='RosterPosition',normalize='all')
```

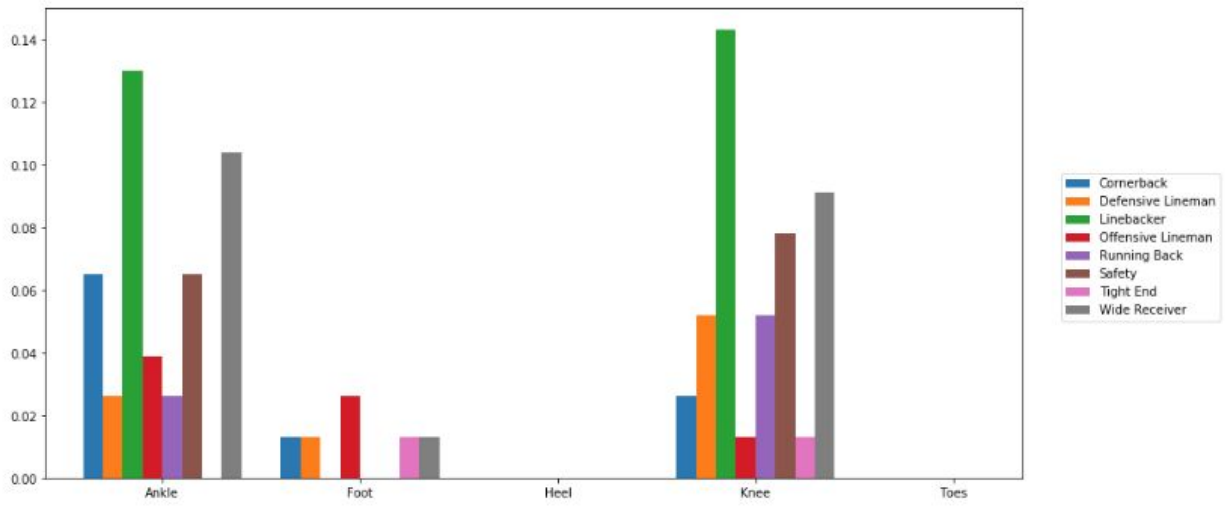


Fig 15: When normalized by number of players in each group, knee injuries afflict running backs more, feet injuries afflict tight ends more and ankle injuries afflict cornerbacks.

```
if (choice=='Knee') or (choice=='Foot') or (choice=='Ankle'):
    dxp.aggplot(agg='BodyPart', data=injuriesplays0, hue='RosterPosition',normalize='RosterPo
sition')
```

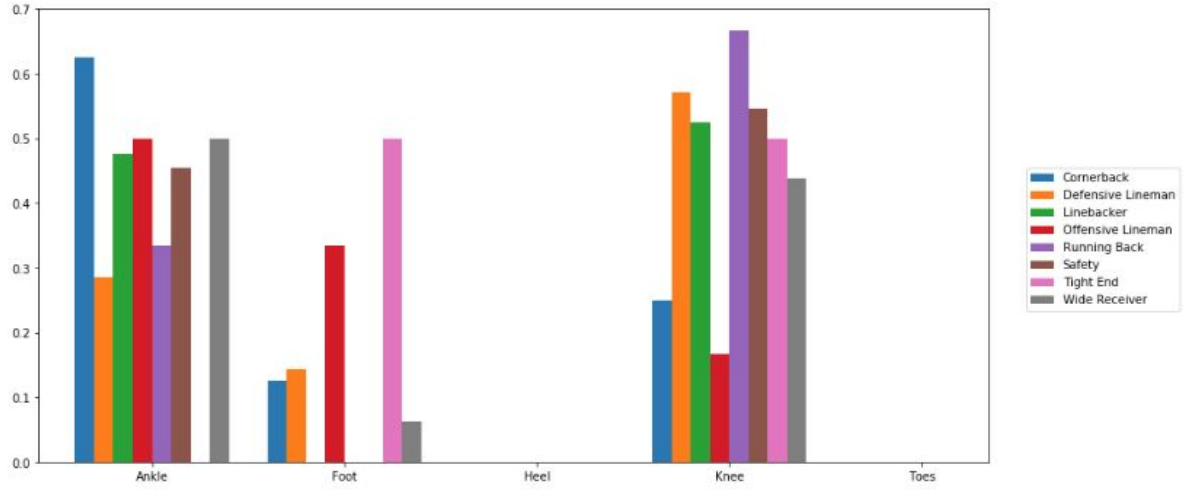


Fig 16: Pass plays across ankle, feet and knees were the source of majority of the injuries.

```
if (choice=='Knee') or (choice=='Foot') or (choice=='Ankle'):
    dxp.aggplot(agg='BodyPart', data=injuriesplays0, hue='PlayType',normalize='BodyPart')
```

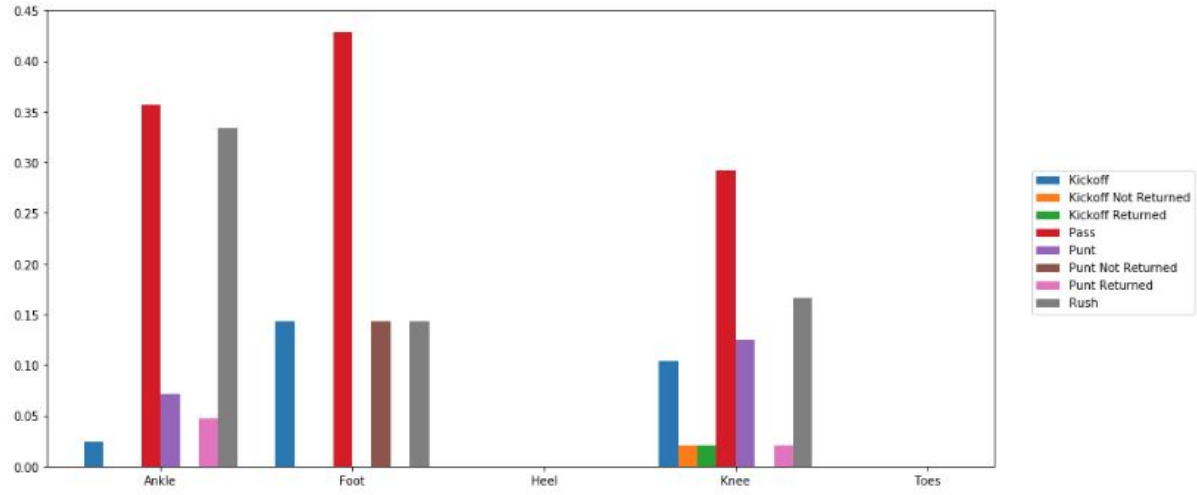


Fig 17: Across all injuries, pass plays were the majority source of injuries.

```
if (choice=='Knee') or (choice=='Foot') or (choice=='Ankle'):
    dxp.aggplot(agg='BodyPart', data=injuriesplays0, hue='PlayType',normalize='all')
```

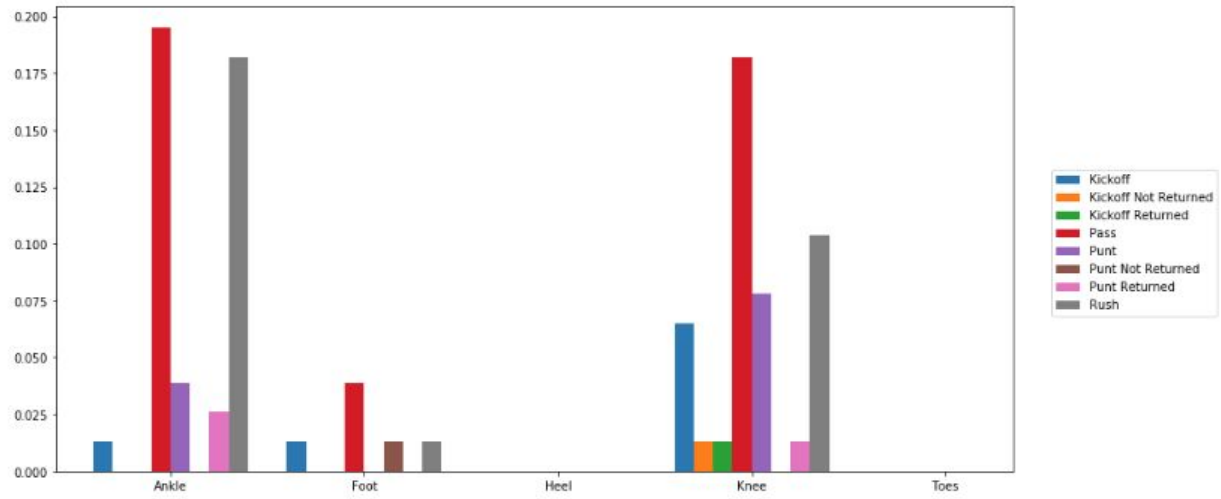


Fig 18: When normalized by playtype, we see a higher prevalence of knee injuries for kickoffs returned/not returned, punts not returned for feet and punt returned for ankles.

```
if (choice=='Knee') or (choice=='Foot') or (choice=='Ankle'):
    dxp.aggplot(agg='BodyPart', data=injuriesplays0, hue='PlayType',normalize='PlayType')
```

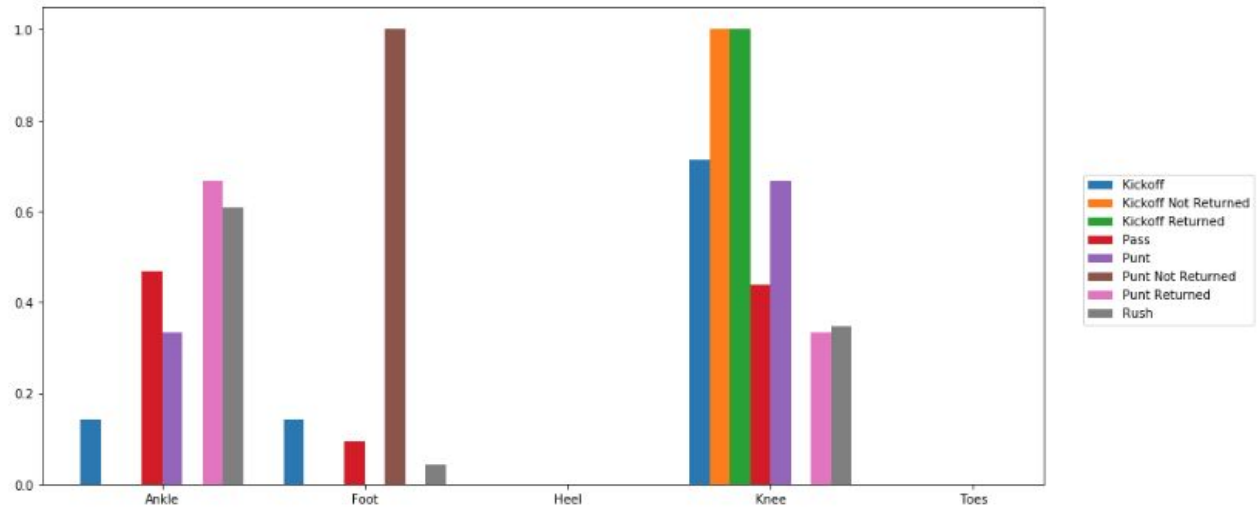


Fig 19: Clear weather is the dominant weather in ankle and knee injuries while cloudy is dominant in feet injuries.

```
if (choice=='Knee') or (choice=='Foot') or (choice=='Ankle'):
    dxp.aggplot(agg='BodyPart', data=injuriesplays0, hue='Weather0',normalize='BodyPart')
```

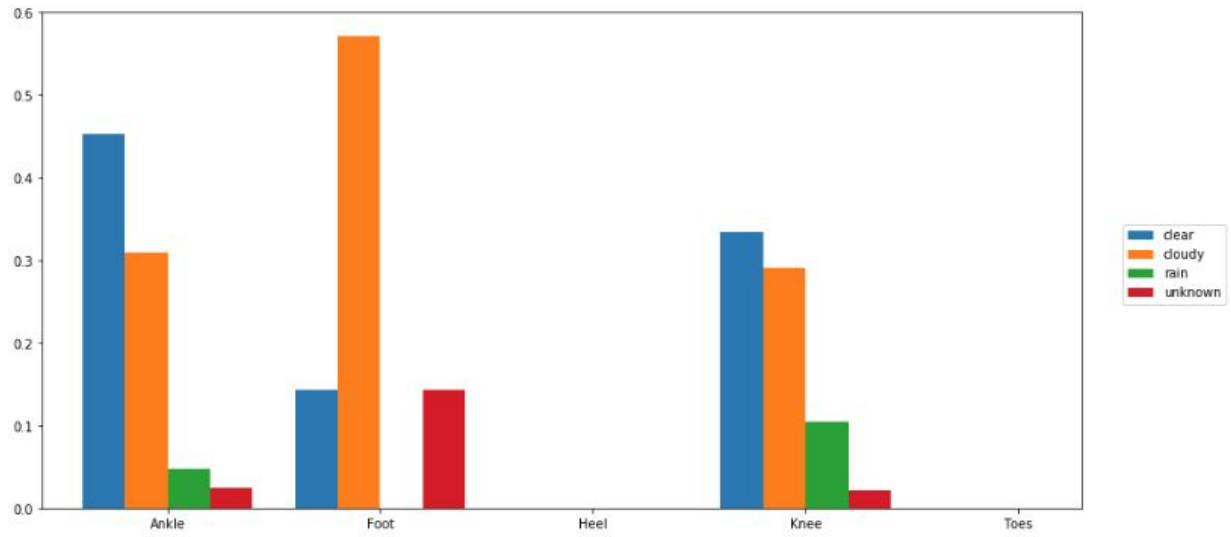


Fig 20: Across all injuries, clear weather is dominant except for feet.

```
if (choice=='Knee') or (choice=='Foot') or (choice=='Ankle'):
    dxp.aggplot(agg='BodyPart', data=injuriesplays0, hue='Weather0',normalize='all')
```

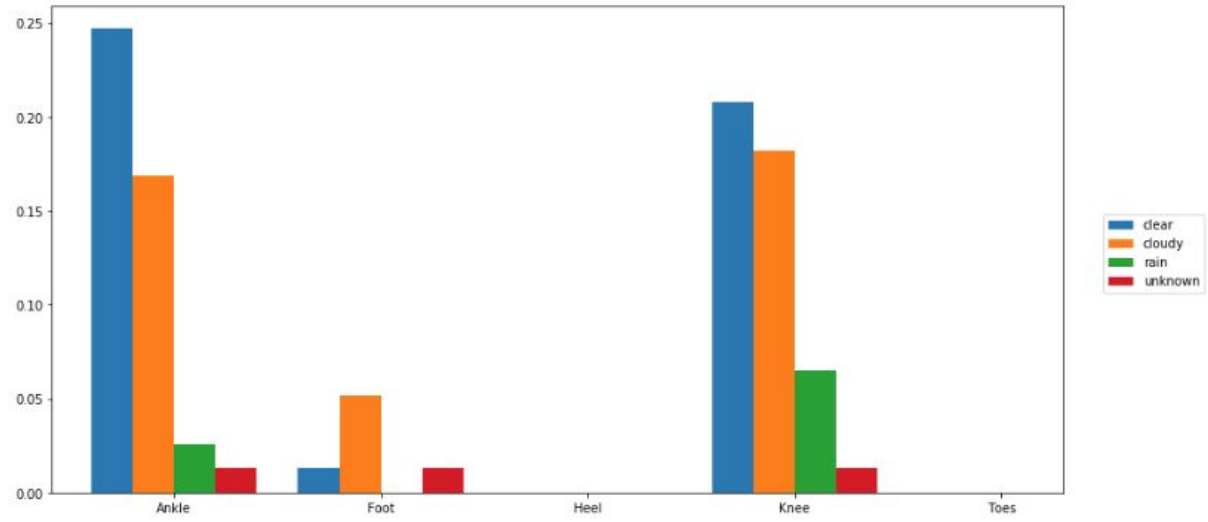


Fig 21: When normalized by weather, we see that rainy weather has a higher prevalence for knee injuries and clear weather for ankle injuries.

```
if (choice=='Knee') or (choice=='Foot') or (choice=='Ankle'):
    dxp.aggplot(agg='BodyPart', data=injuriesplays0, hue='Weather0',normalize='Weather0')
```

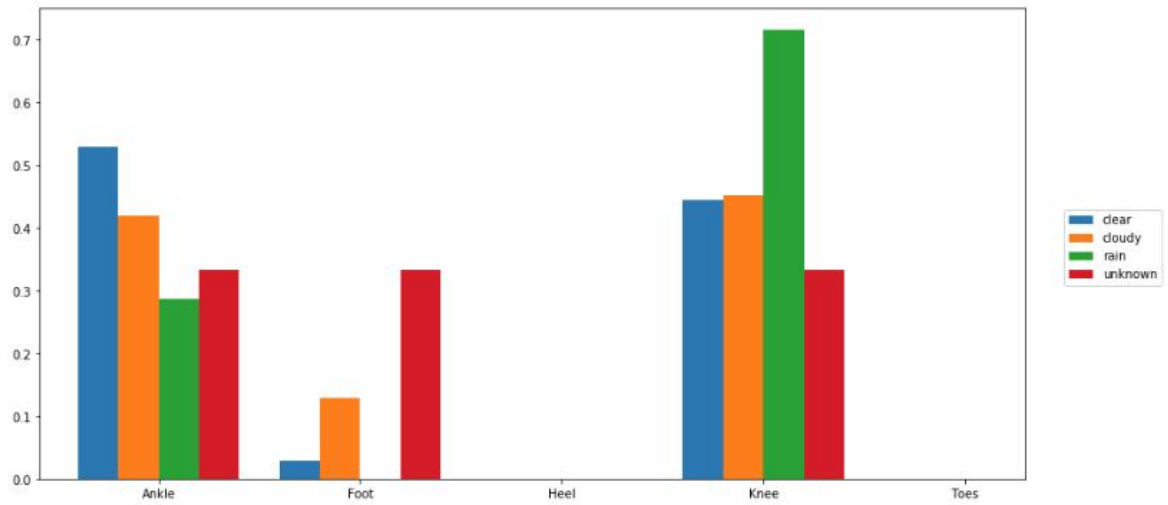
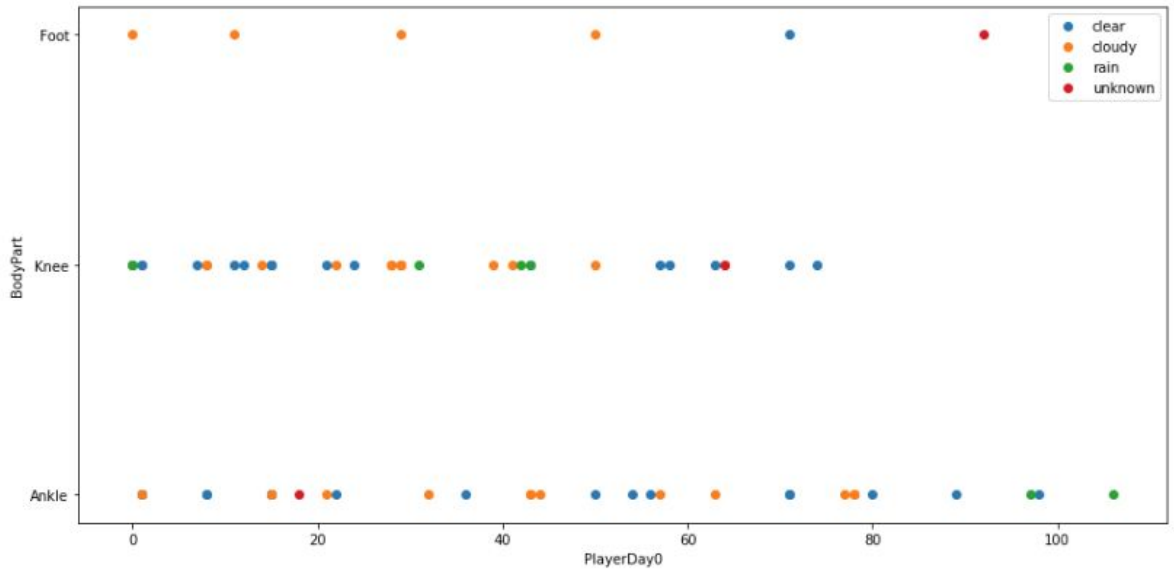


Fig 22: Knee injuries taper off by day 75 in the season.

```
if (choice=='Knee') or (choice=='Foot') or (choice=='Ankle'):
    #dxp.jointplot('experience', 'salary', data=emp, hue='gender')
 dxp.jointplot(x='PlayerDay0', y='BodyPart',hue='Weather0', data=injplayplayers.loc[injplaypl
ayers.Preseason==0,:].drop_duplicates(subset=['PlayKey']))
```



2. KNEE INJURIES

The scores for our three classification models targeting knee injuries were:

CatBoost:

Baseline: Log-loss: 0.0225, Precision: 0.8070, Recall: 0.9988, Fscore: 0.8798

Altered : Log-loss: 0.0204, Precision: 0.8466, Recall: 0.9944, Fscore: 0.9074

Confusion Matrix: [26547, 46]

[1, 104]

LightGBM:

Baseline: Log-loss: 0.0044, Precision: 0.8893, Recall: 0.9852, Fscore: 0.9319

Altered : Log-loss: 0.0044, Precision: 0.8893, Recall: 0.9852, Fscore: 0.9319

Confusion Matrix: [26564, 29]

[3, 102]

Random Forest:

Baseline: Precision: 0.9808, Recall: 0.7857, Fscore: 0.8582

Altered : Precision: 0.9797, Recall: 0.8679, Fscore: 0.9163

Confusion Matrix: [53160, 5]

[99, 132]

Fig 23: Feature Importance:

1		feature	cbnumber	rfnumber	lgbmnumber
2	0	PlayerDay0	1	1	6
3	1	PlayerGame	2	2	1
4	3	Weather	3	4	8
5	2	Position	4	3	3
6	7	StadiumTyp e	5	8	4
7	6	PositionGr oup	6	7	12
8	4	RosterPosi tion	7	5	2
9	5	Temperatur e0	8	6	10
10	9	PlayerGame Play	9	10	11
11	8	FieldType	10	9	7
12	12	IsSunny	11	13	28
13	16	PlayType	12	17	16
14	10	IsCloudy	13	11	5
15	11	IsWet	14	12	27

Fig 24: Feature Interaction:

1		feature1	feature2	relativeimportance	f1name	f2name
2	0	21.0	39.0	3.816911948703478	PlayerGame	PlayerDay0
3	1	20.0	24.0	3.737192675121509	RosterPosition	Weather
4	2	22.0	24.0	2.900605979616815	StadiumType	Weather
5	3	20.0	22.0	2.553168166774968	RosterPosition	StadiumType
6	4	24.0	39.0	2.238766700059333	Weather	PlayerDay0
7	5	21.0	27.0	2.211346888729075	PlayerGame	Position
8	6	20.0	27.0	2.082579660484505	RosterPosition	Position
9	7	24.0	27.0	2.000580720698534	Weather	Position
10	8	21.0	24.0	1.999828255708158	PlayerGame	Weather
11	9	20.0	39.0	1.931232431377189	RosterPosition	PlayerDay0
12	10	22.0	27.0	1.892169696287218	StadiumType	Position
13	11	24.0	28.0	1.838810764892604	Weather	PositionGroup
14	12	22.0	39.0	1.723278230433432	StadiumType	PlayerDay0
15	13	20.0	21.0	1.510521429914358	RosterPosition	PlayerGame

Recursive Feature Elimination:

Features kept: ['x', 'y', 'dis', 'o', 'v_horizontal', 'x_vdiff_adiff_f2delta', 'y_vdiff_adiff_f2delta', 'RosterPosition', 'PlayerGame', 'StadiumType', 'FieldType', 'Weather', 'PlayType', 'PlayerGamePlay', 'Position', 'PositionGroup', 'IsWet', 'IsSunny', 'IsIndoor', 'PlayerDay0', 'Temperature0']

Features eliminated: ['dir', 's', 'v_vertical', 'x_vdiff_f1delta', 'y_vdiff_f1delta', 'dis_vdiff_f1delta', 'dis_vdiff_adiff_f2delta', 'dir_vdiff_f1delta', 'dir_vdiff_adiff_f2delta', 'o_vdiff_f1delta', 'o_vdiff_adiff_f2delta', 's_vdiff_f1delta', 's_vdiff_adiff_f2delta', 'IsSnow', 'IsCloudy', 'IsControlled', 'IsStadiumUnknown', 'IsDomeOpen', 'IsDomeClosed', 'IsOutdoor', 'Preseason']

3. ANKLE INJURIES

The scores for our three classification models targeting ankle injuries were:

CatBoost:

Baseline: Log-loss: 0.0197, Precision: 0.7670, Recall: 0.9943, Fscore: 0.8460

Altered : Log-loss: 0.0130, Precision: 0.8290, Recall: 0.9921, Fscore: 0.8942

Confusion Matrix: [26459, 89]

[2, 156]

LightGBM:

Baseline: Log-loss: 0.0069, Precision: 0.8584, Recall: 0.9799, Fscore: 0.9102

Altered : Log-loss: 0.0067, Precision: 0.8584, Recall: 0.9799, Fscore: 0.9102

Confusion Matrix: [26480, 60]

[6, 152]

Random Forest:

Baseline: Precision: 0.9664, Recall: 0.7718, Fscore: 0.8431

Altered : Precision: 0.9664, Recall: 0.7718, Fscore: 0.8431

Confusion Matrix: [53056, 12]

[146, 174]

Fig 25: Feature Importance:

1		feature	cbnumber	rfnumber	lgbmnumber
2	0	PlayerGame	1	1	1
3	2	Weather	2	3	9
4	4	PlayerDay0	3	5	7
5	1	Position	4	2	5
6	3	RosterPosition	5	4	2
7	6	Temperature0	6	7	10
8	8	StadiumType	7	9	3
9	5	PositionGroup	8	6	6
10	7	FieldType	9	8	4
11	10	IsSunny	10	11	17
12	12	PlayerGamePlay	11	13	11
13	9	IsCloudy	12	10	8
14	20	PlayType	13	21	30
15	13	IsIndoor	14	14	15

Fig 26: Feature Interaction:

1		feature1	feature2	relativeimportance	f1name	f2name
2	0	22.0	27.0	3.5491367791919184	StadiumType	Position
3	1	22.0	24.0	3.544994460931758	StadiumType	Weather
4	2	24.0	27.0	3.3002006337514467	Weather	Position
5	3	20.0	24.0	3.041176440038975	RosterPosition	Weather
6	4	21.0	39.0	2.9593283014373073	PlayerGame	PlayerDay0
7	5	21.0	24.0	2.774940828889478	PlayerGame	Weather
8	6	21.0	27.0	2.658550346291475	PlayerGame	Position
9	7	20.0	27.0	2.30616724932756	RosterPosition	Position
10	8	24.0	39.0	1.8733196362269862	Weather	PlayerDay0
11	9	21.0	22.0	1.5974365313649752	PlayerGame	StadiumType
12	10	20.0	21.0	1.5020136401497242	RosterPosition	PlayerGame
13	11	22.0	39.0	1.4951007245321088	StadiumType	PlayerDay0
14	12	20.0	22.0	1.4536092520173718	RosterPosition	StadiumType
15	13	27.0	39.0	1.2473117872953483	Position	PlayerDay0

Recursive Feature Elimination:

Features kept: ['x', 'y', 'dir', 'dis', 'o', 's', 'v_horizontal', 'v_vertical', 'x_vdiff_f1delta', 'x_vdiff_adiff_f2delta', 'y_vdiff_f1delta', 'y_vdiff_adiff_f2delta', 'dir_vdiff_f1delta', 'dir_vdiff_adiff_f2delta', 'o_vdiff_f1delta', 'o_vdiff_adiff_f2delta', 's_vdiff_adiff_f2delta', 'RosterPosition', 'PlayerGame', 'StadiumType', 'FieldType', 'Weather', 'PlayType', 'PlayerGamePlay', 'Position', 'PositionGroup', 'IsWet', 'IsSunny', 'IsCloudy', 'IsControlled', 'IsIndoor', 'IsOutdoor', 'PlayerDay0', 'Preseason', 'Temperature0']

Features eliminated: ['dis_vdiff_f1delta', 'dis_vdiff_adiff_f2delta', 's_vdiff_f1delta', 'IsSnow', 'IsStadiumUnknown', 'IsDomeOpen', 'IsDomeClosed']

4. FOOT INJURIES

The scores for our three classification models targeting foot injuries were:

CatBoost:

Baseline: Log-loss: 0.0008, Precision: 0.8690, Recall: 0.9998, Fscore: 0.9246

Altered : Log-loss: 0.0005, Precision: 0.9697, Recall: 0.9999, Fscore: 0.9844

Confusion Matrix: [26665, 2]

[0, 31]

LightGBM:

Baseline: Log-loss: 0.0678, Precision: 0.5917, Recall: 0.6123, Fscore: 0.6009

Altered : Log-loss: 0.0677, Precision: 0.5917, Recall: 0.6123, Fscore: 0.6009

Confusion Matrix: [26636, 31]

[24, 7]

Random Forest:

Baseline: Precision: 0.9999, Recall: 0.9796, Fscore: 0.9896

Altered : Precision: 0.9999, Recall: 0.9796, Fscore: 0.9896

Confusion Matrix: [53347, 0]

[2, 47]

Fig 27: Feature Importance:

1		feature	cbnumber	rfnumber	lgbmnumber
2	0	PlayerGame	1	1	10
3	5	Temperature0	2	6	11
4	2	PositionGroup	3	3	33
5	3	Weather	4	4	28
6	1	Position	5	2	15
7	6	PlayerDay0	6	7	6
8	10	StadiumType	7	11	8
9	4	RosterPosition	8	5	20
10	9	FieldType	9	10	40
11	7	IsSunny	10	8	32
12	13	IsOutdoor	11	14	26
13	8	IsCloudy	12	9	42
14	11	IsIndoor	13	12	5
15	15	IsStadiumUnknown	14	16	35

Fig 28: Feature Interaction:

1		feature1	feature2	relativeimportance	f1name	f2name
2	0	20.0	24.0	6.037327266649783	RosterPosition	Weather
3	1	24.0	28.0	3.8092829850398053	Weather	PositionGroup
4	2	24.0	27.0	3.6338241095316 2	Weather	Position
5	3	22.0	24.0	3.2343350952159 26	StadiumType	Weather
6	4	21.0	24.0	3.0658984944387084	PlayerGame	Weather
7	5	20.0	27.0	2.9390013374493256	RosterPosition	Position
8	6	20.0	28.0	2.8347988679163403	RosterPosition	PositionGroup
9	7	22.0	27.0	2.2465345761892976	StadiumType	Position
10	8	21.0	27.0	2.1153570513707325	PlayerGame	Position
11	9	21.0	28.0	2.0974363590076 1	PlayerGame	PositionGroup
12	10	20.0	21.0	2.0183571850606 84	RosterPosition	PlayerGame
13	11	22.0	28.0	1.9092438004025314	StadiumType	PositionGroup
14	12	20.0	22.0	1.8997468221454057	RosterPosition	StadiumType
15	13	21.0	39.0	1.7875235666743186	PlayerGame	PlayerDay0

Recursive Feature Elimination:

Features kept: ['x', 'dis', 's', 'v_horizontal', 'v_vertical', 'x_vdiff_f1delta', 'x_vdiff_adiff_f2delta', 'y_vdiff_f1delta', 'y_vdiff_adiff_f2delta', 'dis_vdiff_f1delta', 'dis_vdiff_adiff_f2delta', 'dir_vdiff_f1delta', 'dir_vdiff_adiff_f2delta', 'o_vdiff_f1delta', 'o_vdiff_adiff_f2delta', 's_vdiff_adiff_f2delta', 'RosterPosition', 'PlayerGame', 'StadiumType', 'FieldType', 'Weather', 'PlayType', 'PlayerGamePlay', 'Position', 'PositionGroup', 'IsWet', 'IsSunny', 'IsIndoor', 'PlayerDay0', 'Temperature0']

Features eliminated: ['y', 'dir', 'o', 's_vdiff_f1delta', 'IsSnow', 'IsCloudy', 'IsControlled', 'IsStadiumUnknown', 'IsDomeOpen', 'IsDomeClosed', 'IsOutdoor', 'Preseason']

5. TOE INJURIES

The scores for our three classification models targeting toe injuries were:

CatBoost:

Baseline: Log-loss: 0.0006, Precision: 0.8939, Recall: 0.9999, Fscore: 0.9406

Altered : Log-loss: 0.0001, Precision: 1.0000, Recall: 1.0000, Fscore: 1.0000

Confusion Matrix: [26672, 0]

[0, 26]

LightGBM:

Baseline: Log-loss: 0.0036, Precision: 0.8939, Recall: 0.9999, Fscore: 0.9406

Altered : Log-loss: 0.0036, Precision: 0.8939, Recall: 0.9999, Fscore: 0.9406

Confusion Matrix: [26665, 7]

[0, 26]

Random Forest:

Baseline: Precision: 1.0000, Recall: 1.0000, Fscore: 1.0000

Altered : Precision: 1.0000, Recall: 1.0000, Fscore: 1.0000

Confusion Matrix: [53353, 0]

[0, 43]

Fig 29: Feature Importance:

1		feature	cbnumber	rfnumber	lgbmnumber
2	1	PlayerGame	1	2	1
3	10	PlayerDay0	2	11	8
4	4	RosterPosition	3	5	7
5	2	Weather	4	3	15
6	5	Position	5	6	3
7	13	StadiumType	6	14	4
8	8	PositionGroup	7	9	14
9	11	Temperature0	8	12	18
10	7	IsIndoor	9	8	10
11	3	FieldType	10	4	6
12	0	IsOutdoor	11	1	2
13	9	IsCloudy	12	10	5
14	12	IsSunny	13	13	24
15	6	Preseason	14	7	9

Fig 30: Feature Interaction:

1		feature1	feature2	relativeimportance	f1name	f2name
2	0	21.0	39.0	4.973739880287815	PlayerGame	PlayerDay0
3	1	20.0	24.0	3.3936653322305372	RosterPosition	Weather
4	2	21.0	28.0	2.8898633735181622	PlayerGame	PositionGroup
5	3	20.0	28.0	2.8407976046444934	RosterPosition	PositionGroup
6	4	21.0	24.0	2.6689792331865023	PlayerGame	Weather
7	5	21.0	27.0	2.5565992844266767	PlayerGame	Position
8	6	22.0	24.0	2.503422760459176	StadiumType	Weather
9	7	20.0	22.0	2.389482519017189	RosterPosition	StadiumType
10	8	24.0	28.0	2.2118850386606286	Weather	PositionGroup
11	9	20.0	21.0	2.203271140311097	RosterPosition	PlayerGame
12	10	20.0	39.0	2.030182505352071	RosterPosition	PlayerDay0
13	11	24.0	39.0	1.8075351243452977	Weather	PlayerDay0
14	12	27.0	28.0	1.7966312916553362	Position	PositionGroup
15	13	22.0	27.0	1.7846316789201504	StadiumType	Position

Recursive Feature Elimination:

Features kept: ['x', 'y', 'dis', 'o', 's', 'v_horizontal', 'v_vertical', 'x_vdiff_f1delta', 'x_vdiff_adiff_f2delta', 'y_vdiff_f1delta', 'y_vdiff_adiff_f2delta', 'dis_vdiff_f1delta', 'dir_vdiff_f1delta', 'dir_vdiff_adiff_f2delta', 'o_vdiff_f1delta', 'o_vdiff_adiff_f2delta', 's_vdiff_f1delta', 'RosterPosition', 'PlayerGame', 'StadiumType', 'FieldType', 'Weather', 'PlayerGamePlay', 'Position', 'PositionGroup', 'IsWet', 'IsSunny', 'IsCloudy', 'IsControlled', 'IsStadiumUnknown', 'IsDomeOpen', 'IsDomeClosed', 'IsOutdoor', 'PlayerDay0', 'Preseason', 'Temperature0']

Features eliminated: ['dir', 'dis_vdiff_adiff_f2delta', 's_vdiff_adiff_f2delta', 'PlayType', 'IsSnow', 'IsIndoor']

6. HEEL INJURIES

The scores for our three classification models targeting heel injuries were:

CatBoost:

Baseline: Log-loss: 0.0001, Precision: 0.7500, Recall: 0.9999, Fscore: 0.8333

Altered : Log-loss: 0.0001, Precision: 1.0000, Recall: 1.0000, Fscore: 1.0000

Confusion Matrix: [26697, 0]

[0, 1]

LightGBM:

Baseline: Log-loss: 0.0996, Precision: 0.5063, Recall: 0.9985, Fscore: 0.5118

Altered : Log-loss: 0.0996, Precision: 0.5063, Recall: 0.9985, Fscore: 0.5118

Confusion Matrix: [26619, 78]

[0, 1]

Random Forest:

Baseline: Precision: 1.0000, Recall: 1.0000, Fscore: 1.0000

Altered : Precision: 1.0000, Recall: 1.0000, Fscore: 1.0000

Confusion Matrix: [53392, 0]

[0, 4]

Fig 31: Feature Importance:

1		feature	cbnumber	rfnumber	lgbmnumber
2	13	PlayType	1	14	16
3	2	Weather	2	3	12
4	6	PlayerGame	3	7	6
5	0	Position	4	1	1
6	8	Temperature0	5	9	7
7	4	PositionGroup	6	5	4
8	17	StadiumType	7	18	5
9	11	PlayerGamePlay	8	12	24
10	1	IsSunny	9	2	20
11	28	x_vdiff_adiff_f2delta	10	29	28
12	3	RosterPosition	11	4	2
13	10	PlayerDay0	12	11	10
14	9	y	13	10	40
15	5	FieldType	14	6	35

Fig 32: Feature Interaction:

1		feature1	feature2	relativeimportance	f1name	f2name
2	0	24.0	25.0	2.5428130530544513	Weather	PlayType
3	1	20.0	27.0	2.297729566726383	RosterPosition	Position
4	2	25.0	31.0	2.085267992640834	PlayType	IsSnow
5	3	24.0	27.0	1.950455682082855	Weather	Position
6	4	24.0	28.0	1.8850338980811479	Weather	PositionGroup
7	5	25.0	28.0	1.7855577881264808	PlayType	PositionGroup
8	6	27.0	31.0	1.7070121499350115	Position	IsSnow
9	7	20.0	25.0	1.6675065748234412	RosterPosition	PlayType
10	8	22.0	24.0	1.6545338769377809	StadiumType	Weather
11	9	22.0	25.0	1.5294056235331717	StadiumType	PlayType
12	10	20.0	31.0	1.5275627635967473	RosterPosition	IsSnow
13	11	25.0	27.0	1.5162282622353755	PlayType	Position
14	12	20.0	28.0	1.5137423054704755	RosterPosition	PositionGroup
15	13	1.0	31.0	1.4263284053178205	y	IsSnow

Recursive Feature Elimination:

Features kept: ['x_vdiff_f1delta', 'x_vdiff_adiff_f2delta', 'y_vdiff_f1delta', 'y_vdiff_adiff_f2delta', 'dis_vdiff_f1delta', 'dis_vdiff_adiff_f2delta', 'dir_vdiff_f1delta', 'dir_vdiff_adiff_f2delta', 'o_vdiff_f1delta', 'o_vdiff_adiff_f2delta', 's_vdiff_f1delta', 's_vdiff_adiff_f2delta', 'RosterPosition', 'PlayerGame', 'StadiumType', 'FieldType', 'Weather', 'PlayType', 'PlayerGamePlay', 'Position', 'PositionGroup', 'IsWet', 'IsSunny', 'IsSnow', 'IsCloudy', 'IsControlled', 'IsStadiumUnknown', 'IsDomeOpen', 'IsDomeClosed', 'IsIndoor', 'IsOutdoor', 'PlayerDay0', 'Preseason', 'Temperature0']

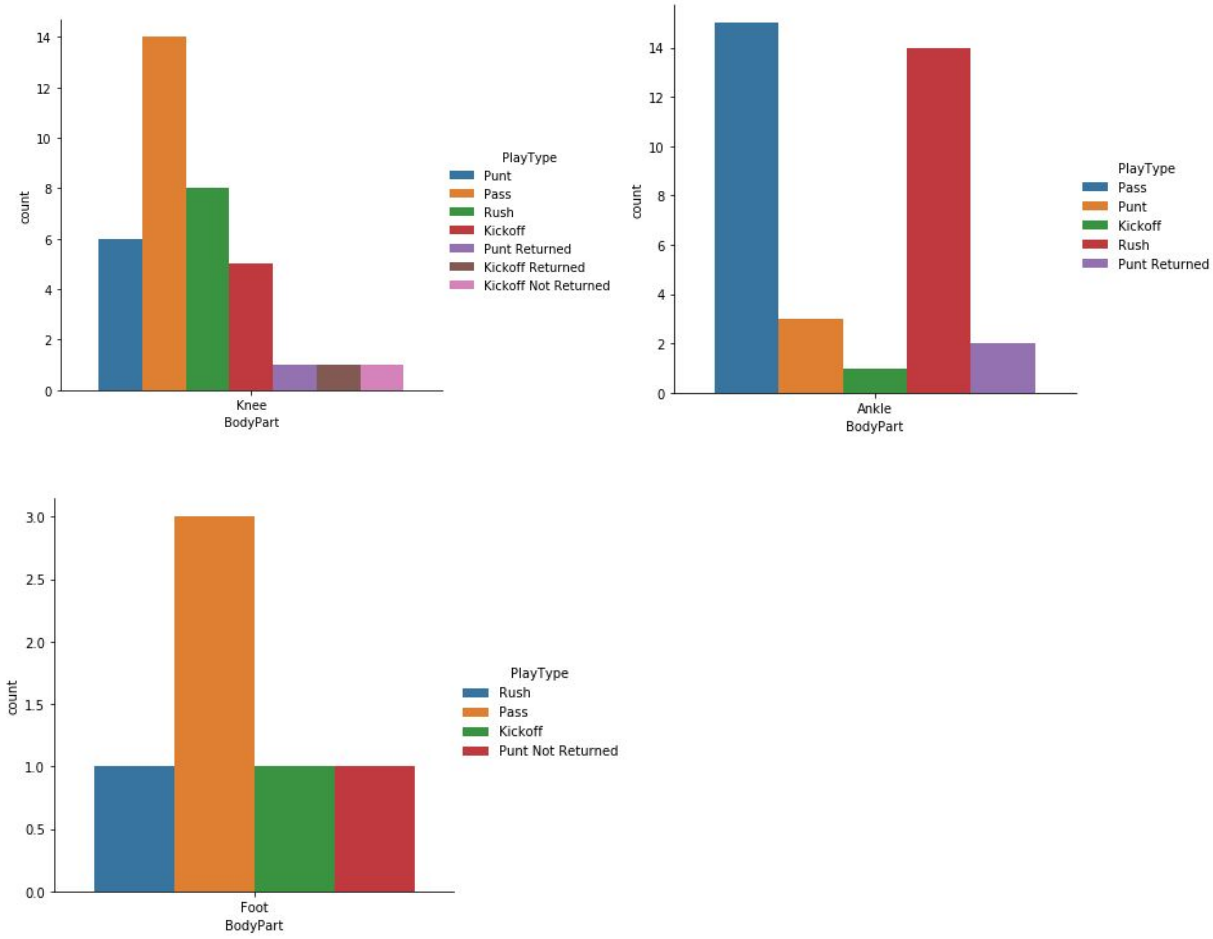
Features eliminated: ['x', 'y', 'dir', 'dis', 'o', 's', 'v_horizontal', 'v_vertical']

DISCUSSION AND CONCLUSION

1. **Special teams:**

The figures below suggest that special plays (Punt, Punt Returned, Kickoff, Kickoff not returned) have a low prevalence for injury. However, it should be noted that most plays are passes and rushes, and our analysis might not be representative of the imbalance in the dataset.

Fig 33:

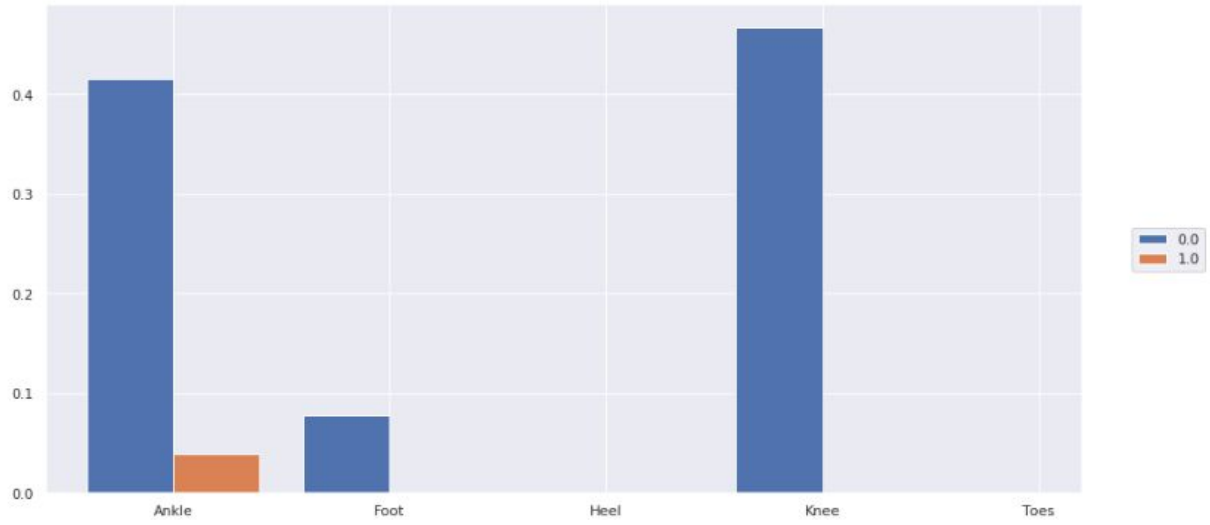


2. **Injuries during practice/preseason:**

While we have data on preseason games, we have no data on practice sessions. From the figure below, we can see that there were only ankle injuries in preseason, and they formed a minority.

Fig 34:

```
if (choice=='Knee') or (choice=='Foot') or (choice=='Ankle'):
    dxp.aggplot(agg='BodyPart', data=injuriesplays0, hue='Preseason',normalize='all')
```



3. **Backups more prone to injury:**

There was not enough data to make a distinction between starters and substitutes.

4. **Coaching decisions:**

Given the occurrence or reinjuries, this may suggest that the coaching staff is not instituting a proper rotation, or the recovery regime is not allowing enough time for injuries to heal.

5. **Coming back from injury/% healthy:**

In the table below, there is a sample of 4 players that got reinjured within the same season.

PlayerKey 33337 injured his foot in an earlier game 2 and reinjured it in game 8. PlayerKey 43540 injured his ankle in game 3 and reinjured it in game 7. PlayerKey 45950 injured his toes in game 6 and went on to injure his ankle in game 8. PlayerKey 44449 injured his knee in game 6 and went on to injure his ankle in game 28.

Given the small dataset, we cannot conclude outside of PlayerKey 43540 and 33337, that one type of injury is casual to the next injury or establish a relationship.

The data might be suggesting that the current recovery time regime might not be enough leeway for the injury to heal in full.

Fig 35:

	Player Key	GameID	PlayKey	Body Part	Surface	DM_M1	DM_M7	DM_M28	DM_M42	RecoveryTime
46	33337	33337-2		Foot	Natural	1	1	1	1	42
85	33337	33337-8	33337-8-15	Foot	Natural	1	1	1	0	28

35	43540	43540-3	43540-3-14	Ankle	Natural	1	0	0	0	1
44	43540	43540-7	43540-7-2	Ankle	Natural	1	1	1	1	42
37	44449	44449-6	44449-6-13	Knee	Natural	1	0	0	0	1
77	44449	44449-28	44449-28-35	Ankle	Synthetic	1	0	0	0	1
34	45950	45950-8	45950-8-18	Ankle	Natural	1	0	0	0	1
49	45950	45950-6		Toes	Synthetic	1	1	0	0	7

6. **Injured in cold temperatures/more start/stops:**

Comparing the overall distribution of temperature to the temperature at which injuries occurred, we see a gulf between 0 – 35 degrees where there is almost a non-existent occurrence of injuries.

Fig 36:

```
dxp.jointplot('PlayerDay0','Temperature', data=playersplay.loc[playersplay['Temperature']>0], hue='Surf
```

```
:
<matplotlib.axes._subplots.AxesSubplot at 0x7f96dc2b3b00>
```

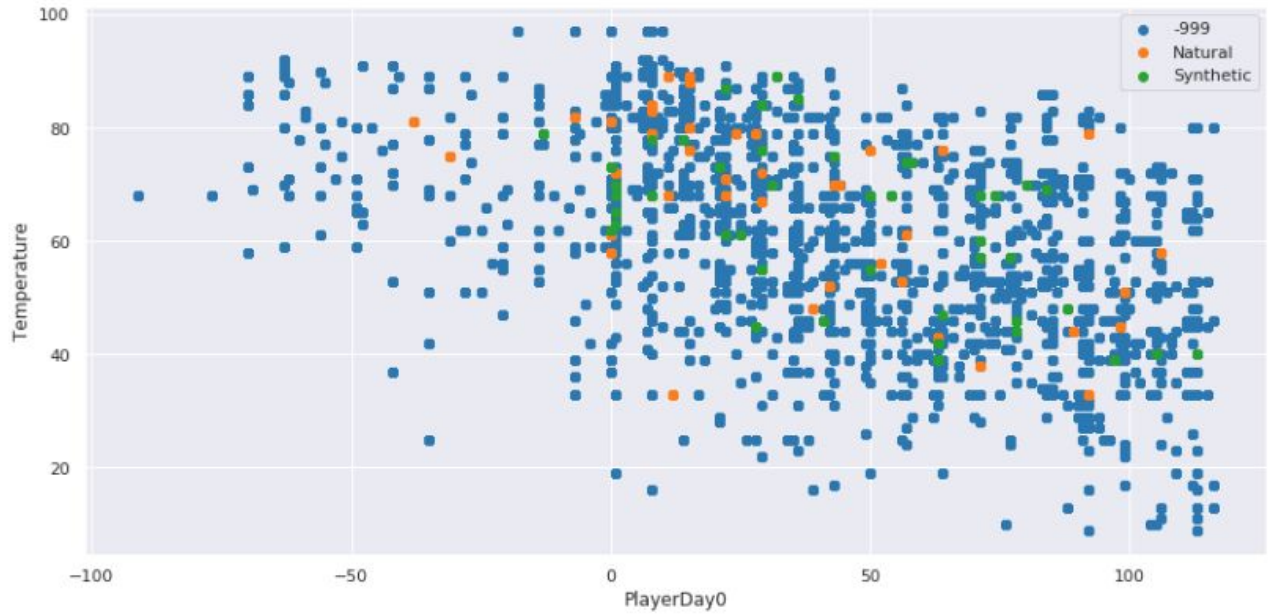
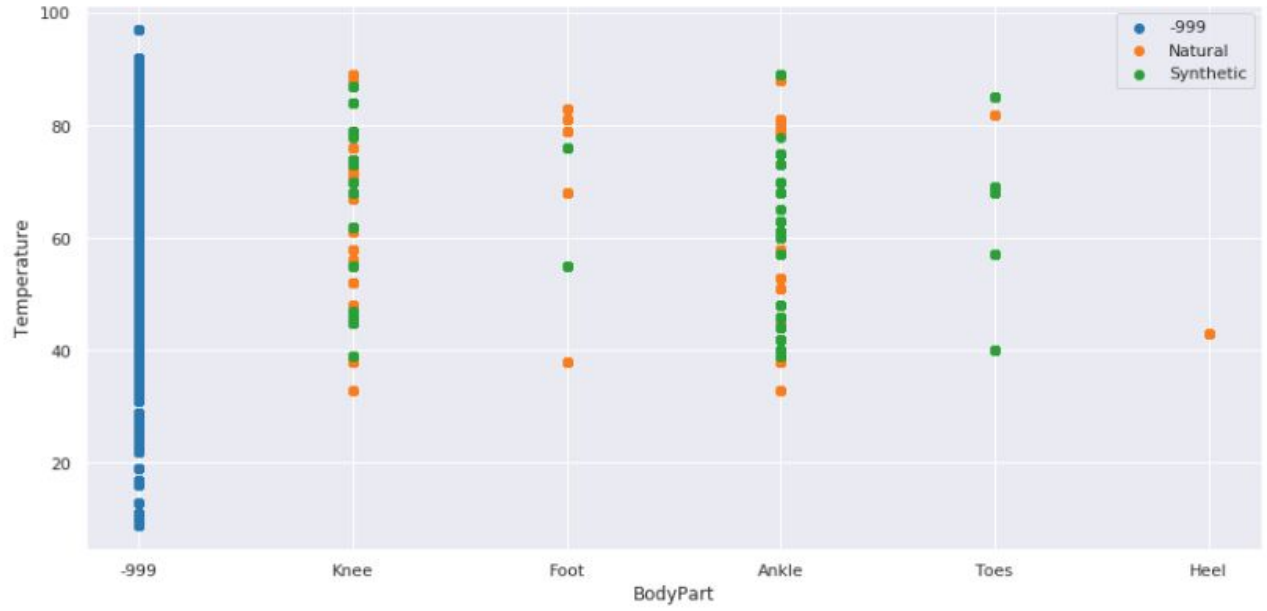


Fig 37:

```
dxp.jointplot('BodyPart','Temperature', data=playersplay.loc[playersplay['Temperature']>0], hue='Surfac
]:
<matplotlib.axes._subplots.AxesSubplot at 0x7f96dc365128>
```



7. **Same injury, more severe recovery time on synthetic:**

In the figure below, outside of foot injuries, injuries sustained on synthetic turf (brown) either equal or exceed those on natural turf by count. Moving across the different groups to the more severe recovery times, synthetic turf injuries outnumber natural turf injuries.

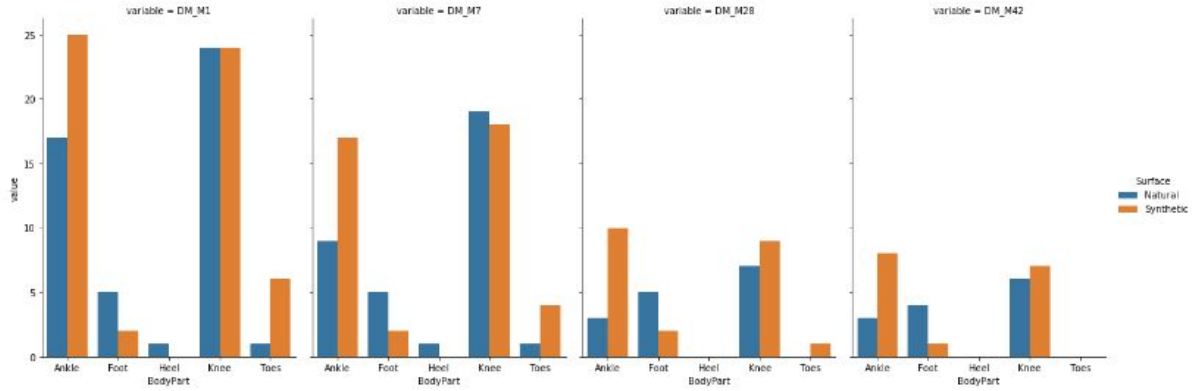
This hypothesis needs more data as our current dataset is too small to be conclusive.

Fig 38:

```
Inj_Surf_BodyPart_New = pd.melt(Inj_Surf_BodyPart, id_vars=['Surface','BodyPart'], value_vars
=['DM_M1','DM_M7','DM_M28','DM_M42'])

sns.catplot(x='BodyPart', y='value', hue='Surface', col='variable',
            data=Inj_Surf_BodyPart_New, kind='bar', height=6, aspect=.7)
```

```
<seaborn.axisgrid.FacetGrid at 0x7f8a27aeb780>
```



8. **Less injuries during rain/snow:**

Fig 39: There were no injuries during snow weather

```
if (choice=='Knee') or (choice=='Foot') or (choice=='Ankle'):
    dxp.aggplot(agg='BodyPart', data=injuriesplays0, hue='IsSnow',normalize='all')
```

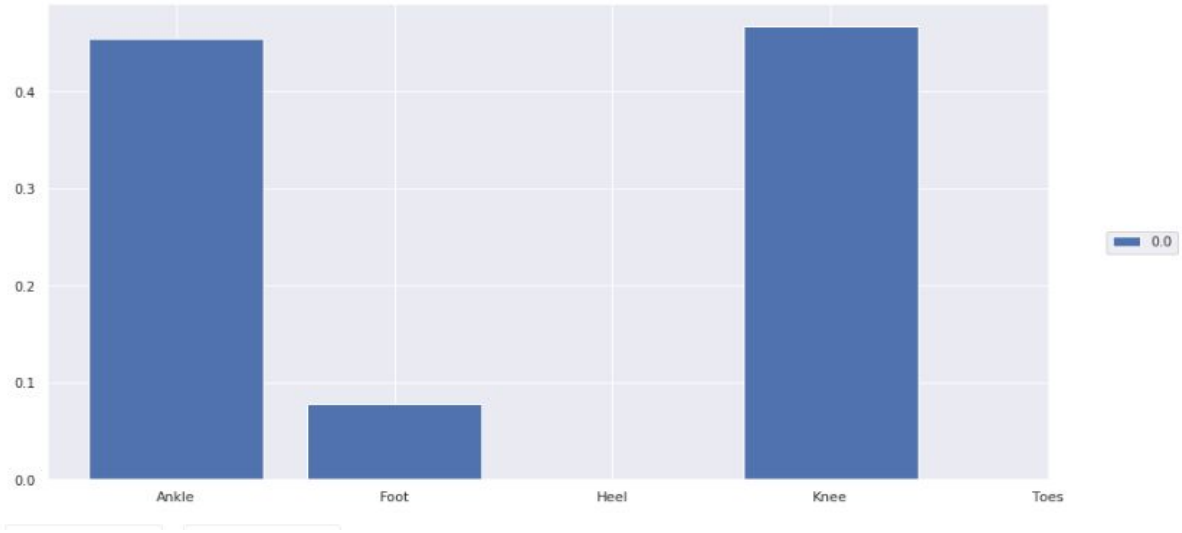


Fig 40: There is a very low occurrence of injuries during wet weather (rain or snow).

```
if (choice=='Knee') or (choice=='Foot') or (choice=='Ankle'):
    dxp.aggplot(agg='BodyPart', data=injuriesplays0, hue='IsWet',normalize='all')
```

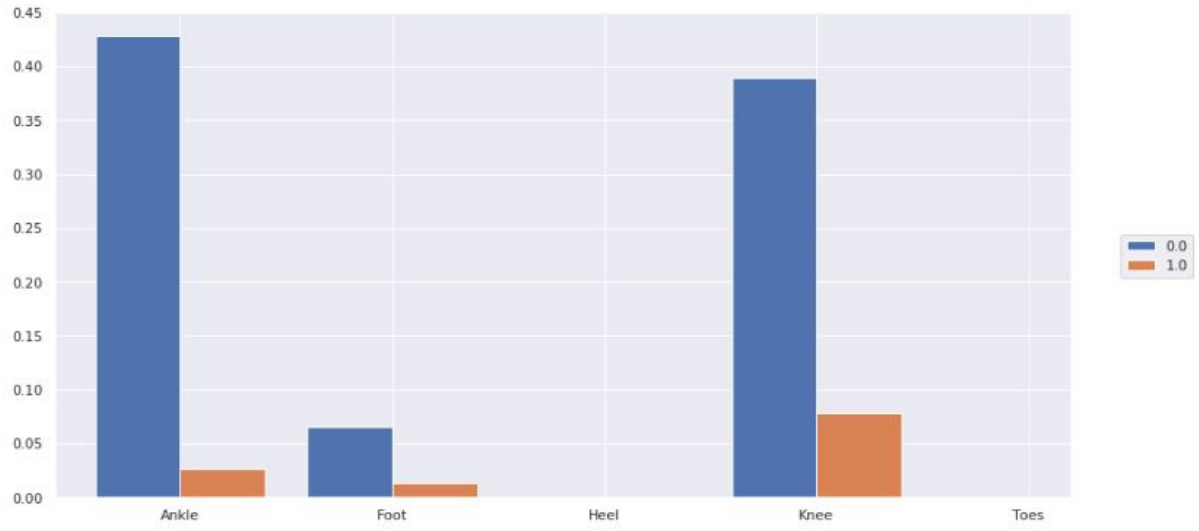
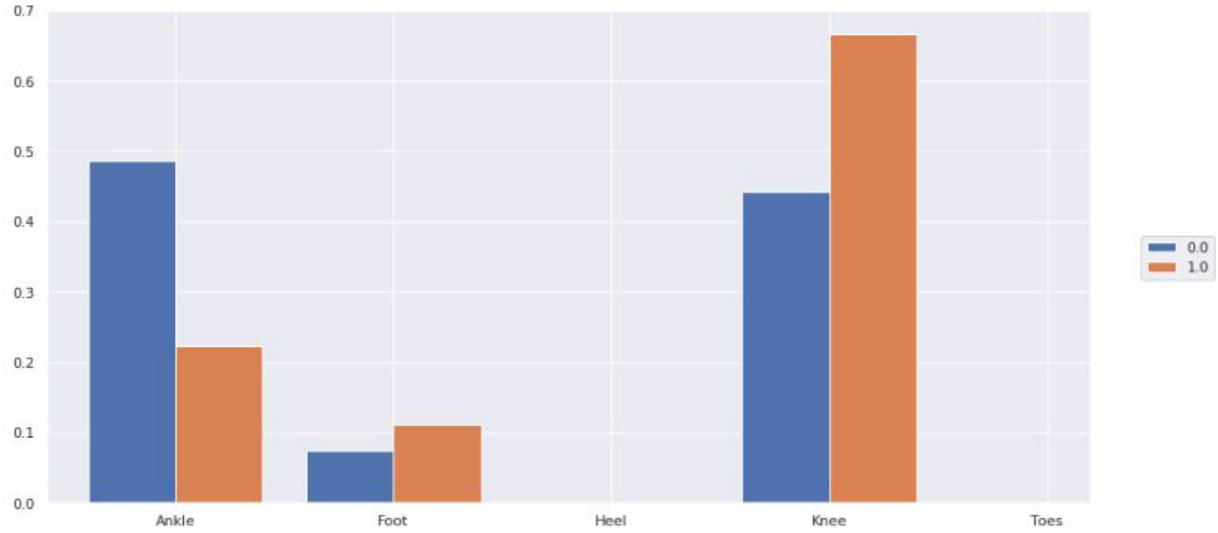


Fig 41: However, when normalized by IsWet, we can see that there was a higher prevalence of injuries during wet weather when considering the few samples of wet weather.

```
if (choice=='Knee') or (choice=='Foot') or (choice=='Ankle'):
    dxp.aggplot(agg='BodyPart', data=injuriesplays0, hue='IsWet',normalize='IsWet')
```



9. Games with multiple injuries:

Given our hypothesis, we would assume given the perfect set of conditions prevalent to injuries, then particular games would present opportunities for more injuries. The analysis

discovered only one game with more than one injury, however this injury occurred to the same player during the same play.

Fig 42:

	PlayerKey	GameID	PlayKey	BodyPart	Surface	DM_M1	DM_M7	DM_M28	DM_M42	Recovery Time
86	47307	47307-10	47307-10-18	Knee	Synthetic	1	1	0	0	7
87	47307	47307-10	47307-10-18	Ankle	Synthetic	1	1	0	0	7

The conclusion is that the injury dataset is too small to investigate this hypothesis

REFERENCES

Angular/Linear speed i.e.
https://www.s-cool.co.uk/a-level/physics/circular-motion/revise-it/angles-in-radians-and-angular-speed-versus-linear-speed